**DOCTOR OF SCIENCES**

# Study of block diagonal preconditioners using partial spectral information to solve linear systems arising in constrained optimization problems

Tannier, Charlotte

*Award date:*
2016

*Awarding institution:*
University of Namur


**UNIVERSITÉ DE NAMUR**

FACULTÉ DES SCIENCES

DÉPARTEMENT DE MATHÉMATIQUE

# Study of block diagonal preconditioners using partial spectral information to solve linear systems arising in constrained optimization problems

Thèse présentée par
**Charlotte Tannier**
pour l'obtention du grade
de Docteur en Sciences

Composition du Jury:

Anders Forsgren
Anne Lemaitre (Président du Jury)
Dominique Orban
Daniel Ruiz
Annick Sartenaer (Promoteur)

2016

Imprimé en Belgique

ISBN : 978-2-87037-950-9
Dépôt légal: D / 2016/ 1881/ 33

Université de Namur
Faculté des Sciences
rue de Bruxelles, 61, B-5000 Namur (Belgique)

**Étude de préconditionneurs bloc diagonaux utilisant de l'information spectrale partielle pour résoudre des systèmes linéaires dans les problèmes d'optimisation avec contraintes**

par Charlotte Tannier

**Résumé :** Ce travail a pour objectif le développement et l'étude de nouveaux préconditionneurs bloc diagonaux pour résoudre des systèmes linéaires indéfinis ayant une forme de point-selle. Nous considérons le préconditionneur bloc diagonal « idéal » proposé par Murphy, Golub et Wathen (2000) basé sur le complément de Schur exact, et nous nous concentrons sur le cas où le bloc (1,1) admet seulement quelques très petites valeurs propres. En supposant que l'information exacte sur ces valeurs propres et ces vecteurs propres associés est disponible, nous proposons différentes approximations du préconditionneur bloc diagonal de Murphy, Golub et Wathen et nous analysons les propriétés spectrales des matrices préconditionnées. Nous généralisons les résultats théoriques aux systèmes découlant des méthodes de points intérieurs et nous illustrons numériquement la performance des préconditionneurs proposés. Enfin, nous analysons l'intéraction entre les blocs (1,1) et (1,2) des systèmes de point-selle et nous étudions les situations dans lesquelles les petites valeurs propres du bloc (1,1) peuvent avoir un impact sur la convergence des méthodes itératives.

**Study of block diagonal preconditioners using partial spectral information to solve linear systems arising in constrained optimization problems**

by Charlotte Tannier

**Abstract:** This work is concerned with the development and the study of novel block diagonal preconditioners for solving indefinite linear systems with a saddle - point form. We consider the « ideal » block diagonal preconditioner proposed by Murphy, Golub and Wathen (2000) based on the exact Schur complement, and we focus on the case where the (1,1) block has few very small eigenvalues. Assuming that the exact information on these eigenvalues and their associated eigenvectors is available, we propose different approximations of the block diagonal preconditioner of Murphy, Golub and Wathen and we analyse the spectral properties of the preconditioned matrices. We generalize the theoretical results on systems arising in interior-point methods and we illustrate the performance of the proposed preconditioners through some numerical experiments. We finally analyse the interaction between the (1,1) and (1,2) blocks of saddle-point systems and we study the circumstances in which small eigenvalues of the (1,1) block can have an impact on the convergence of iterative methods.

Thèse de doctorat en Sciences Mathématiques (Ph.D. thesis in Mathematics)
Date: 11/07/2016
Département de Mathématique
Promoteur (Advisor): Prof. A. Sartenaer

# Remerciements

Il est naturel de remercier à la fin d'un tel travail tous ceux qui, plus ou moins directement, ont contribué à le rendre possible. C'est avec un enthousiasme certain que je profite de ces quelques lignes pour rendre hommage aux personnes qui ont participé à leur manière à la réalisation de cette thèse.

En premier lieu, je tiens à remercier ma promotrice, Annick Sartenaer, pour la confiance qu'elle m'a accordée en acceptant d'encadrer ce travail doctoral. Merci pour ta franchise, merci de m'avoir posé un bon nombre de fois La bonne question (commençant généralement par « Pourquoi ...? »), bien que souvent angoissante, cela m'a permis de me remettre en question et de me dépasser. Ta relecture finale méticuleuse de chacun de mes chapitres a sans aucun doute contribué à améliorer considérablement la qualité et la clarté de ce manuscrit. Merci encore pour nos nombreuses discussions qui dépassent de loin le sujet de cette thèse.

Je souhaite aussi remercier sincèrement Daniel Ruiz pour avoir fait germer en moi l'envie de réaliser cette thèse de doctorat au travers de mon stage de master. Ce « coup de foudre scientifique » a été pour moi, le début d'une belle et longue aventure. Je le remercie également pour nos nombreux échanges de vues et nos moments passés « au tableau » qui vont évidemment me manquer. Sans tes qualités scientifiques et humaines, cette période de ma vie n'aurait pas été aussi enrichissante et agréable.

Mes remerciements vont également aux autres membres du jury : Anders Forsgren, Anne Lemaitre, présidente du jury et Dominique Orban pour leur bienveillance, leurs remarques pertinentes et constructives lors de la défense privée de ma thèse.

Merci à tous les membres du département. J'ai trouvé ici un nouveau chez moi et j'ai l'impression d'y avoir passé une partie essentielle de ma vie. Merci aux anciens dont j'ai eu la chance de croiser la route : Julien, Patrick, Nicolas, Jehan, Charlotte, Sandrine, Charles, Jérémy. En particulier, merci aux incorruptibles du matin de la salle café : Eric, Anne, André H., Benoit. L'endroit idéal pour bien commencer la journée. Merci à toi André F. pour tes passages au bureau, tes encouragements et ta compréhension. Merci à vous Pascale et Alice pour vos conseils et votre aide au quotidien. Merci à vous Eve et Delphine pour vos passages réguliers au bureau et votre bonne humeur.

Merci aux « jeunes » : Manon, Marie M., Julien, Morgane, Marie P., Jon, Pauline. Merci pour ces moments de franche rigolade à l'arsenal et ces débats toujours des plus passionnants. Merci à toi Martine, tu as toujours eu les mots justes. Merci à toi Audrey, pour toutes nos conversations et nos confidences.

Nous avons commencé ensemble mais tu as fini avant moi, merci à toi Emilie. Même à distance (pas tant que ça :-) ), tu m'écoutes, me comprends, me fais rire depuis longtemps maintenant, partages mes joies et mes peines. Sans une amie comme toi à mes côtés, mon optimisme aurait bien souvent disparu. A mes yeux aussi, notre entente a toujours été parfaite.

Merci à vous les Anne-So. Sans vous, ces dernières années n'auraient pas été aussi agréables. Merci Anne-Sophie C. pour ta complicité quotidienne. Je garderai en tête la façon unique que tu as de me faire rire, ton bureau rangé à la perfection, ton caractère bien trempé mais surtout, ta volonté de m'aider même quand tu es débordée et ta compréhension lors des moments difficiles. Merci Anne-Sophie L. pour tes bonjours matinaux (qui vont me manquer), pour tes conseils toujours justes, pour tes avis tranchés et pour tes innombrables encouragements. Vous allez tellement me manquer les filles !

Enfin je souhaite remercier tous les miens, qui ont fait de moi la personne que je suis aujourd'hui. Il est clairement plus facile d'écrire des mathématiques que de témoigner en quelques lignes de toute l'attention, l'affection et l'amour pour mes proches. Je les remercie pour leur présence, leur confiance et toutes ces choses que ces mots ne disent pas. Je souhaite remercier particulièrement, mes parents qui me soutiennent toujours et m'entourent. Vous êtes mes racines et vous m'avez transmis la persévérance, la volonté et le goût de la vie. Merci aussi à toi marraine et à toi bonne-maman pour votre soutien qui a rendu ces derniers mois un peu moins stressants et m'a permis de gagner quelques heures de sommeil.

Les mots me manquent pour remercier, à sa juste valeur, mon conjoint, Mathieu qui m'a supportée pendant toute la durée de ma thèse et plus particulièrement durant les derniers mois de rédaction qui n'ont pas été faciles. Merci d'avoir accepté que je sois très souvent dans ma bulle, d'avoir été (trop) souvent mon sas de décompression et d'avoir géré seul ces derniers temps nos deux enfants. Cette thèse et moi te devons beaucoup. Merci !

Pour finir, merci à mes deux rayons de soleil, Maxence et Annaline. Vous m'avez souvent obligé à déconnecter, à prendre du recul dans les moments difficiles et vous m'avez montré que le plus important n'est pas écrit dans ces pages.

Encore un grand merci à tous pour m'avoir conduite à ce jour mémorable.

Charlotte

# Contents

# Thesis organization and main contribution

The scope of this work is the construction and the study of new efficient preconditioners for indefinite linear systems arising in constrained optimization. We consider (possibly large and sparse) saddle-point linear systems resulting from the Karush-Kuhn-Tucker optimality conditions. We assume that the (1,1) block, $A$ say, is symmetric and positive definite and possibly ill-conditioned with relatively few very small eigenvalues. We propose two variants for a block diagonal preconditioner based on a Schur complement approximation derived from some prior spectral information extracted from $A$ directly, i.e., using information on the subspace associated to the smallest eigenvalues in $A$. We study the spectral properties of the preconditioned matrix in both cases and illustrate their numerical performance on a standard problem in fluid dynamics. This in turn leads to the study of the interaction between blocks in saddle-point systems and to highlight some aspects of this interaction. Through the Schur complement approximation based on the very small eigenvalues of $A$ and theoretical developments, we analyze how and in which circumstances the ill-conditioning due to these eigenvalues effectively spoils the convergence of iterative methods like Krylov subspace methods.

## Structure of the thesis

The thesis is organized as follows.

**Chapter** 1 introduces several linear systems that one needs to solve in unconstrained and constrained optimization methods. We focus on the indefinite linear systems with a block structure called saddle-point systems. More precisely, we study KKT systems, related to the Karush-Kuhn-Tucker optimality conditions, and symmetric quasi-definite (SQD) systems. We then address the properties associated to both kinds of system. The large range of methods which can solve symmetric indefinite linear systems shows the importance of this topic in the optimization community.

**Chapter** 2 leads the reader in the field of iterative techniques for solving linear systems and develops the framework of preconditioners that will serve as the basis for our contributions. Some methods belonging to the class of Krylov subspace methods are presented. We also discuss the block diagonal preconditioner studied in Golub, Greif and Varah (2006). Most of the results of Chapter 1 and Chapter 2 can be found in the literature. However, these results are adapted and presented in a form suitable to motivate the design and allow the forthcoming analysis of preconditioners.

**Chapter** 3 presents approximations of the inverse of the (1,1) block and of the Schur complement using spectral information extracted from this (1,1) block in KKT systems and SQD systems. We analyse in both cases the eigenvalues distribution of the preconditioned Schur complement and we illustrate on a test example.

**Chapter** 4 proposes two new approximations of the "ideal" block diagonal preconditioner given by Murphy, Golub and Wathen (2000) using the previous approximations of the inverse of the (1,1) block and of the Schur complement. We next study the eigenvalues distribution of preconditioned KKT and SQD systems and we compare these two preconditioners. Finally, we focus on KKT systems to develop in detail the formulation of the preconditioners combined with a first level of preconditioning.

**Chapter** 5 analyses and compares the behaviour of the spectral preconditioners introduced previously on a problem in fluid dynamics called the Stokes problem and generated with the MATLAB package IFISS.

**Chapter** 6 focusses on KKT systems and first analyses the interaction between blocks through the inverse of the Schur complement approximation. We show the configurations according to which the influence of the small eigenvalues of $A$ can have an effect on the convergence of iterative methods. We give a toy example and some numerical illustrations based on the explicit construction of a (1,2) block to give some intuition of this interaction. We next refine the bounds on the eigenvalues of a KKT matrix by theoretical developments and finally give a new cheapest alternative to our preconditioners.

**Chapter** 7 presents some practical considerations. We first analyse how to combine a first level of preconditioning with the preconditioner developed in Giraud, Ruiz and Touhami (2006) which allows to extract partial spectral information of the (1,1) block. A practical approach, showing a practical implementation of the approximation of the inverse of the (1,1) block, is developed in the last part of this chapter.

## Contribution

If a cited result is already available in the literature, we only mention a reference without giving a proof. The main contribution based on work with my supervisor Annick Sartenaer and my colleague Daniel Ruiz, are summarized as follows:

1. Theorem 2.9 (Chapter 2) analyses and motivates the choice of some parameter $\omega$ introduced in the block diagonal preconditioner developed by Golub et al. (2006).

2. We propose in Section 3.2 for KKT matrices and in Section 3.2.2 for SQD matrices spectral approximations of the inverse of the Schur complement using an approximation of the inverse of the (1,1) block known as spectral low rank update approach developed by Carpentieri, Duff and Giraud (2003) . Theorem 3.1 and Theorem 3.4 establish lower and upper bounds on the eigenvalues of the preconditioned Schur complement in both cases.

3. Chapter 4 introduces two spectral preconditioners based on spectral approximations of the inverse of the (1,1) block and of the Schur complement introduced in Chapter 3 for the KKT and SQD systems. Theorems 4.1 and 4.3 for KKT matrices and Theorems 4.4 and 4.5 for SQD matrices analyse their spectral properties. Sections 4.3 and 4.4 give a comparison between both preconditioners. Section 4.5 combines these preconditioners with a first level of preconditioning for the KKT systems.

4. Chapter 5 illustrates the numerical behaviour of the proposed preconditioners on the Stokes problem.

5. Chapter 6 studies in detail the interaction between blocks in KKT systems with a first analysis through the Schur complement approximation in Section 6.1 and some illustrations in Section 6.2. Section 6.3, by means of a theoretical analysis, refines the interval in Rusten and Winther (1992), associated to the positive eigenvalues in KKT matrices (Theorem 6.10). Finally, section 6.4 studies the possibility of reducing the low rank update in the inverse of the approximation of the Schur complement and generalizes the block diagonal preconditioner.

6. Based on the work of Golub, Ruiz and Touhami (2007), Chapter 7 gives new theoretical results (Theorems 7.1 and 7.2) on how to combine a first level preconditioner with the preconditioner developed in Golub et al. (2007) to extract desired spectral information of the (1,1) block. Section 7.2 introduces an approximation of the spectral low rank update approach of the (1,1) block (Theorems 7.4 and 7.5).

A first paper related to this work and entitled "Using partial spectral information for block diagonal preconditioning of saddle-point systems" has been submitted to COAP[(1)]. A second paper based on Chapter 6 is currently in preparation.

[(1)] Computational Optimization and Applications.

Chapter 1

# Introduction to optimization methods leading to systems of equations

The topic of interest of this thesis is a detailed study of large linear systems, called *saddle-point systems*, of the form

$$\begin{bmatrix} A & B \\ B^T & -C \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}, \tag{1.1}$$

where the square matrix $A$ of order $n$ is symmetric, sparse and ill-conditioned, and such that some spectral information on $A$ is available. The matrix $B$ is rectangular of size $n \times m$ with $m \leq n$, while the square matrix $C$ is of order $m$ and symmetric. The matrix in (1.1),

$$\mathcal{A} := \begin{bmatrix} A & B \\ B^T & -C \end{bmatrix}, \tag{1.2}$$

is often called a *saddle-point matrix* or, in the case where $C = 0$, a *KKT matrix*, in reference to the Karush-Kuhn-Tucker's first-order necessary optimality conditions used to solve constrained optimization problems. In the case where the matrices $A$ and $C$ are positive definite, the matrix $\mathcal{A}$ is called *symmetric quasi-definite*, or *SQD* for short. We focus on the solution of saddle-point linear systems of the KKT form,

$$\mathcal{A}_{KKT} x = b \equiv \begin{bmatrix} A & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}, \tag{1.3}$$

where $A \in \mathbb{R}^{n \times n}$ is symmetric and $B \in \mathbb{R}^{n \times m}$ has a full column rank ($m \leq n$) and on systems of the SQD form,

$$\mathcal{A}_{SQD}x = b \equiv \begin{bmatrix} A & B \\ B^T & -C \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}, \tag{1.4}$$

where $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{n\times n}$ are symmetric positive definite and $B \in \mathbb{R}^{n\times m}$ $(m \leq n)$.

The system (1.1) arises in many areas of computational science and engineering, as in computational fluid dynamics (Elman, Silvester and Wathen, 2005), in electromagnetism (Perugia, Simoncini and Arioli, 1999), in optimal control (Battermann and Heinkenschloss, 1997, Battermann and Sachs, 2001) and in weighted least-squares problems (Björck, 1996). This system also emerges as subproblems in different methods for general constrained optimization (Gill, Murray and Wright, 1981, Nocedal and Wright, 2006). In particular, the SQD matrix appears in regularized interior-point methods (Friedlander and Orban, 2012). Since the late 1990s, there has been a surge of interest in saddle-point systems. Hence, numerous solution techniques have been proposed for this type of systems. Benzi, Golub and Liesen (2005) gave a first survey in which they presented a set of methods for solving these systems.

This work cannot reasonably cover all the different areas involving the solution of saddle-point systems. Instead, we focus on the KKT and SQD systems and we motivate the topic of this thesis by giving the general context of some optimization results in the unconstrained and constrained cases. Section 1.1 introduces the general concepts of unconstrained optimization that will be needed in the constrained case. Section 1.2 presents classes of constrained optimization methods leading to the solution of KKT or SQD systems. Finally, we establish that the KKT and SQD matrices are indefinite and we also introduce important properties of these matrices such as their invertibility or the existence of a factorization.

We only focus in this chapter on tools that we need afterwards and we highlight the systems that are of interest in the remainder of this work. This overview of unconstrained and constrained optimization methods is thus far from complete. We refer the reader to the books of Gill et al. (1981) and Nocedal and Wright (2006) for more details.

## 1.1 Unconstrained optimization

A general unconstrained optimization problem is defined as

$$\min_{x\in\mathbb{R}^n} f(x), \tag{1.5}$$

where the *objective function* $f : \mathbb{R}^n \to \mathbb{R}$ denotes the (continuous) function to minimize. A *global minimizer* of this problem is a vector $x^\star \in \mathbb{R}^n$ satisfying

$$f(x^\star) \leq f(x) \quad \forall x \in \mathbb{R}^n.$$

Generally, it can be difficult to find a global minimizer of $f$ while the identification of a point $x^\star$ achieving the smallest value of $f$ in a neighborhood $\mathcal{N}$ of $x^\star$ is easier. Such a point is called a *local minimizer* and this terminology distinguishes it from a *strict local minimizer*, which is a point $x^\star$ such that $f(x^\star) < f(x)$ for all $x \in \mathcal{N}$ with $x \neq x^\star$. When the objective function $f$ is convex, we have a strong property given by the following theorem.

**Theorem 1.1** When $f$ is a convex function, any local minimizer $x^\star$ is a global minimizer of $f$. If in addition $f$ is differentiable, then any stationary point $x^\star$ is a global minimizer of $f$.

**Proof.** See, e.g., (Nocedal and Wright, 2006, p.16) □

Assuming that the objective function $f$ is at least twice continuously differentiable ($f \in \mathcal{C}^2$), we define the gradient and the Hessian of $f$.

**Definition 1.1** The *gradient of* $f$ is the vector of first partial derivatives whose $i$-th component is $\partial f(x)/\partial x_i$, and is denoted by $\nabla f(x)$.

**Definition 1.2** The *Hessian of* $f$ is the matrix of second partial derivatives whose $i, j$-th component is $\partial^2 f(x)/\partial x_i \partial x_j$, and is denoted by $\nabla^2 f(x)$.

### 1.1.1 Optimality conditions

The following theorem states the first and the second-order necessary optimality conditions for problem (1.5). Assuming that $x^\star$ is a local minimizer, one can deduce properties about the gradient $\nabla f(x^\star)$ and the Hessian $\nabla^2 f(x^\star)$.

**Theorem 1.2**
**First-order necessary optimality conditions**
If $x^\star$ is a local minimizer of $f$ and $f$ is continuously differentiable in an open neighborhood of $x^\star$, then $\nabla f(x^\star) = 0$.
**Second-order necessary optimality conditions**
If $x^\star$ is a local minimizer of $f$ and $\nabla^2 f$ exists and is continuous in an open neighborhood of $x^\star$, then $\nabla f(x^\star) = 0$ and $\nabla^2 f(x^\star)$ is positive semidefinite.

**Proof.** See, e.g., (Nocedal and Wright, 2006, pp.14 – 15) □

We can guarantee that $x^\star$ is a strict local minimizer of $f$ if we have sufficient conditions on the derivatives of $f$ at the point $x^\star$.

**Theorem 1.3 Second-order sufficient optimality conditions**
If $\nabla^2 f$ is continuous in an open neighborhood of $x^\star$, $\nabla f(x^\star) = 0$ and $\nabla^2 f(x^\star)$ is positive definite, then $x^\star$ is a strict local minimizer of $f$.

**Proof.** See, e.g., (Nocedal and Wright, 2006, p.16) □

We now consider a well-known problem in unconstrained optimization, the case where the objective function is quadratic. Let us suppose that

$$q(x) := \frac{1}{2} x^T A x - b^T x, \tag{1.6}$$

with $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$. By the second-order sufficient optimality conditions, if $x^\star$ is a local minimizer, the gradient given by $\nabla q(x^\star) = Ax^\star - b$ has to be zero and the Hessian, $\nabla^2 q(x^\star) = A$, has to be positive definite. In this case, the minimizer $x^\star$ of $q(x)$ is the unique solution of a linear system of the form

$$Ax = b, \tag{1.7}$$

with a symmetric positive definite matrix $A$. Solving this kind of linear system is an active research area and is very important in scientific computing. We will come back to this in Chapter 2.

### 1.1.2 Newton's method

In practice, at the light of the first-order necessary optimality condition, practical algorithms rely on the (approximate) solution of the nonlinear system

$$\nabla f(x) = 0$$

of $n$ equations with $n$ unknowns, to solve the optimization problem (1.5). We consider *Newton's method* which is an iterative method used to generate a sequence of points, starting at an initial guess $x_0$ and hopefully converging towards a point $x^\star$ at which $\nabla f(x^\star) = 0$ holds. To compute the next iterate $x_{k+1}$ from the current iterate $x_k$, the objective function is replaced by a second-order Taylor approximation built around this current iterate $x_k$,

$$f(x_k + p) \approx f(x_k) + \nabla f(x_k)^T p + \frac{1}{2} p^T \nabla^2 f(x_k) p,$$

where $p \in \mathbb{R}^n$. This quadratic approximation of $f$ around $x_k$ has a unique minimizer if $\nabla^2 f(x_k)$ is positive definite. Its gradient,

$$\nabla f(x_k) + \nabla^2 f(x_k) p,$$

is then set equal to zero, which leads to a set of linear equations to solve,

$$\nabla^2 f(x_k)p = -\nabla f(x_k), \tag{1.8}$$

which are known as the *Newton equations*. If the Hessian $\nabla^2 f(x_k)$ is positive definite, the solution $p_k$ is thus unique and one can define the next iterate as

$$x_{k+1} = x_k + p_k.$$

Note that Newton's method often does not converge if the initial point $x_0$ is not close enough to a local minimizer. The line-search and trust-region approaches (see, e.g., Conn, Gould and Toint, 2000 and Nocedal and Wright, 2006) are two strategies which ensure, under appropriate assumptions, global convergence, i.e., convergence from any starting point. Note also that in many cases, the Hessian can be indefinite or singular, or unavailable or too expensive to compute at every iteration; then Newton's method cannot be applied. In practice, one uses quasi-Newton methods that do not need to directly evaluate the Hessian but that use a suitable approximation (see, e.g., Nocedal and Wright, 2006).

Observe that (1.8) is a symmetric system. In constrained optimization, Newton's method is also used in some approaches but leading to symmetric systems of equations of the KKT or SQD form. In Section 1.2, some methods for constrained optimization problems are considered and we highlight these systems.

## 1.2 Constrained optimization

We now add equality and inequality constraints to (1.5)

$$\begin{array}{ll} \min\limits_{x\in\mathbb{R}^n} & f(x) \\ \text{s.t.} & \left\{ \begin{array}{ll} c_i(x) = 0, & i \in \mathcal{E}, \\ c_i(x) \geq 0, & i \in \mathcal{I}, \end{array} \right. \end{array} \tag{1.9}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is the objective function and $c_i : \mathbb{R}^n \to \mathbb{R}$ for all $i \in \mathcal{E} \cup \mathcal{I}$ are the *constraint functions*. The two disjoint index sets $\mathcal{E}$ and $\mathcal{I}$ correspond to the indices of the equality and inequality constraints respectively, and are of size $n_\mathcal{E}$ and $n_\mathcal{I}$. For later use, we also introduce the functions $c_\mathcal{E} : \mathbb{R}^n \to \mathbb{R}^{n_\mathcal{E}}$ and $c_\mathcal{I} : \mathbb{R}^n \to \mathbb{R}^{n_\mathcal{I}}$ defined as $c_\mathcal{E}(x)^T = [c_1(x) \dots c_{n_\mathcal{E}}(x)]$ and $c_\mathcal{I}(x)^T = [c_1(x) \dots c_{n_\mathcal{I}}(x)]$, respectively.

**Definition 1.3** The *feasible set* $\Omega$ is the set of *feasible points* $x$, i.e., those that satisfy the constraints,

$$\Omega := \{x \in \mathbb{R}^n \text{ such that } c_i(x) = 0, i \in \mathcal{E} \text{ and } c_i(x) \geq 0, i \in \mathcal{I}\}.$$

Extending the definitions of the unconstrained case, we obtain the following definitions for the different types of solutions of problem (1.9). A *global solution* of (1.9) is a vector $x^\star \in \Omega$ satisfying

$$f(x^\star) \leq f(x) \quad \forall x \in \Omega.$$

A vector $x^\star \in \Omega$ is called a *local solution* when $f(x^\star) \leq f(x)$ forall $x \in \mathcal{N} \cap \Omega$, where $\mathcal{N}$ is a neighborhood of $x^\star$. Finally, we define a *strict local solution* as a point $x^\star \in \Omega$ such that $f(x^\star) < f(x)$ for all $x \in \mathcal{N} \cap \Omega$ with $x \neq x^\star$.

**Definition 1.4** For $c : \mathbb{R}^n \to \mathbb{R}^m$, the *Jacobian* $J(x)$ is defined as the $n \times m$ matrix

$$J(x) := [\nabla c_1(x) \ \nabla c_2(x) \ \ldots \ \nabla c_m(x)]^T,$$

where $c_i$ is the $i$th component of $c$.

We denote by $J_{\mathcal{E}}(x) \in \mathbb{R}^{n_{\mathcal{E}} \times n}$ and $J_{\mathcal{I}}(x) \in \mathbb{R}^{n_{\mathcal{I}} \times n}$ the Jacobian of the equality and inequality constraints, respectively.

### 1.2.1 Optimality conditions

As in the unconstrained case, there exist first and second-order optimality conditions. In this section, we only focus on first-order necessary optimality conditions. We first define some concepts and notation needed in the remainder of this section.

**Definition 1.5** The *Lagrangian function* for problem (1.9) is defined as

$$\mathcal{L}(x, y, z) := f(x) - y^T c_{\mathcal{E}}(x) - z^T c_{\mathcal{I}}(x),$$

where the components of $y \in \mathbb{R}^{n_{\mathcal{E}}}$ and $z \in \mathbb{R}^{n_{\mathcal{I}}}$ are called the *Lagrange multipliers*.

**Definition 1.6** The *active set* $\mathcal{A}(x)$ at a given feasible point $x$ is the set of indices of the constraints $c_i$ satisfied as equalities at $x$,

$$\mathcal{A}(x) := \mathcal{E} \cup \{i \in \mathcal{I} \text{ such that } c_i(x) = 0\}.$$

At a feasible point $x$, any inequality constraint satisfied as equality at $x$ is called *active*, while any inequality constraint satisfied as strict inequality is called *inactive*.

**Definition 1.7** The *linear independence constraint qualification* (LICQ) condition holds at a feasible point $x$ for problem (1.9) if the gradients of the active constraints at $x$,

$$\{\nabla c_i(x), \, i \in \mathcal{A}(x)\},$$

are linearly independent.

The first-order necessary optimality conditions given in the following theorem are called the *Karush-Kuhn-Tucker conditions*, or *KKT conditions* for short, and state first-order necessary optimality conditions for $x^\star$ to be a local solution. These conditions were derived by Karush in his master's thesis at the University of Chicago in 1939, but it is only in 1950, when the mathematicians Kuhn and Tucker published their work, that the theory of constrained optimization emerged (see Karush, 1939, and Kuhn and Tucker, 1950).

**Theorem 1.4**
**First-order necessary optimality conditions**
Suppose that $x^\star$ is a local solution of (1.9), that the functions $f$ and $\{c_i\}_{i\in\mathcal{E}\cup\mathcal{I}}$ in (1.9) are continuously differentiable and that the LICQ condition holds at $x^\star$. Then there are Lagrange multiplier vectors $y^\star \in \mathbb{R}^{n_\mathcal{E}}$ and $z^\star \in \mathbb{R}^{n_\mathcal{I}}$, with components $y_i^\star, i \in \mathcal{E}$, and $z_i^\star, i \in \mathcal{I}$, such that the following conditions are satisfied at $(x^\star, y^\star, z^\star)$

$$\begin{aligned}
\nabla_x \mathcal{L}(x^\star, y^\star, z^\star) &= 0, && (1.10)\\
c_\mathcal{E}(x^\star) &= 0, && (1.11)\\
c_\mathcal{I}(x^\star) &\geq 0, && (1.12)\\
z_i^\star &\geq 0, \text{ for all } i \in \mathcal{I}, && (1.13)\\
z_i^\star c_i(x^\star) &= 0, \text{ for all } i \in \mathcal{I}. && (1.14)
\end{aligned}$$

**Proof.** See, e.g., (Nocedal and Wright, 2006, Section 12.4) □

Condition (1.10) is known as the *stationarity condition*. By Definition 1.5 of the Lagrangian function, we can rewrite the stationarity condition as

$$\nabla_x \mathcal{L}(x^\star, y^\star, z^\star) = \nabla f(x^\star) - J_\mathcal{E}(x^\star)^T y^\star - J_\mathcal{I}(x^\star)^T z^\star = 0,$$

with $J_\mathcal{E}(x^\star)$ and $J_\mathcal{I}(x^\star)$ the Jacobian of the equality and inequality constraints at $x^\star$, respectively. The conditions (1.11) and (1.12) imply the feasibility of $x^\star$, while conditions (1.14) imply that either the inequality constraint $i$ is active or the associated Lagrange multiplier is nul, or possibly both. These conditions are known as *the complementarity conditions*. In addition, condition (1.13) imposes nonnegative Lagrange multipliers for the inequality constraints.

For later use (in Chapter 6, Section 6.3.3), we also give the F. John Theorem below (see, e.g., Hiriart-Urruty, 1996) that states first-order necessary optimality conditions for problem (1.9) without any constraint qualification condition.

**Theorem 1.5** Suppose that $x^\star$ is a local solution of (1.9) and that the functions $f$ and $\{c_i\}_{i\in\mathcal{E}\cup\mathcal{I}}$ in (1.9) are continuously differentiable. Then there is a scalar $u^\star \in \mathbb{R}$ and vectors $y^\star \in \mathbb{R}^{n_\mathcal{E}}$ and $z^\star \in \mathbb{R}^{n_\mathcal{I}}$, with components $y_i^\star, i \in \mathcal{E}$, and $z_i^\star, i \in \mathcal{I}$ such that the vector $[u^\star, y^\star, z^\star] \in \mathbb{R} \times \mathbb{R}^{n_\mathcal{E}} \times \mathbb{R}^{n_\mathcal{I}}$ is nonzero and the following conditions are satisfied at $(x^\star, u^\star, y^\star, z^\star)$ with

$$\begin{aligned} u^\star \nabla f(x^\star) - J_\mathcal{E}(x^\star)^T y^\star - J_\mathcal{I}(x^\star)^T z^\star &= 0, \\ z_i^\star &\geq 0, \text{ for all } i \in \mathcal{I}, \\ z_i^\star c_i(x^\star) &= 0, \text{ for all } i \in \mathcal{I}. \end{aligned}$$

**Proof.** See, e.g., (Hiriart-Urruty, 1996, Theorem 3.1) □

In the next sections, we focus on some methods for constrained optimization leading to the solution of KKT or SQD systems. We first consider the quadratic programming, which often appears as subproblems in some methods for general constrained optimization. Then we focus on some approaches to solve general constrained optimization problems as, for instance, the sequential quadratic programming leading to KKT systems, or the interior-point method leading to SQD systems.

### 1.2.2 Quadratic programming

In this section, we consider the constrained optimization problem where the objective function is quadratic and the constraints are all equalities and linear. This type of problem arises in several applications but also as subproblems in methods for general constrained optimization, such as sequential quadratic programming, augmented Lagrangian methods and interior-point methods. Let us consider the general form

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & q(x) := \frac{1}{2}x^T Ax - b^T x \\ \text{s.t.} \quad & B^T x = c, \end{aligned} \tag{1.15}$$

where $A = A^T \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$, $B^T \in \mathbb{R}^{m\times n}$ is the Jacobian of the constraints ($m \leq n$) and $c \in \mathbb{R}^m$. The Lagrangian function for the quadratic problem (1.15) is given by

$$\mathcal{L}(x, y) = \frac{1}{2}x^T Ax - b^T x - y^T(B^T x - c),$$

where $y \in \mathbb{R}^m$ is the vector of Lagrange multipliers. The KKT conditions (1.10) - (1.11) for problem (1.15) state that there exist vectors $x^\star \in \mathbb{R}^n$ and $y^\star \in \mathbb{R}^m$ such that the following system of equations is satisfied,

$$\begin{cases} Ax^\star - b - By^\star = 0 \\ B^T x^\star - c = 0 \end{cases}$$

or, equivalently,

$$\begin{bmatrix} A & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} x^\star \\ -y^\star \end{bmatrix} = \begin{bmatrix} b \\ c \end{bmatrix}. \tag{1.16}$$

The KKT conditions for problem (1.15) thus amount to solve the KKT system (1.16).

### 1.2.3 Sequential quadratic programming

In this section, we consider the method for constrained problems called *sequential quadratic programming* (SQP) introduced by Wilson (1963). We only focus on the case where we have equality constraints so that the problem can be stated as

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & c_{\mathcal{E}}(x) = 0. \end{aligned} \tag{1.17}$$

The idea behind this method is to solve, at a given iteration $k$, a quadratic subproblem of the form (1.15), and to use its solution to construct the next iterate. One replaces the objective function at the current iterate $x_k \in \mathbb{R}^n$ by its local quadratic approximation defined by a second-order Taylor approximation at $x_k$,

$$f(x) \approx f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2}(x - x_k)^T \nabla^2 f(x_k)(x - x_k),$$

while the constraints are linearized around $x_k$,

$$c_{\mathcal{E}}(x) \approx c_{\mathcal{E}}(x_k) + J_{\mathcal{E}}(x_k)(x - x_k),$$

with $J_{\mathcal{E}}(x)$ denoting the Jacobian of the equality constraints. If we set $p = x - x_k$, we get a quadratic problem of the following form

$$\begin{aligned} \min_{p \in \mathbb{R}^n} \quad & f(x_k) + \nabla f(x_k)^T p + \frac{1}{2} p^T \nabla^2 f(x_k) p \\ \text{s.t.} \quad & c_{\mathcal{E}}(x_k) + J_{\mathcal{E}}(x_k) p = 0. \end{aligned} \tag{1.18}$$

As we have seen in the previous section, we have that the KKT conditions (1.10) - (1.11) are equivalent to solve the following system of equations for problem (1.18),

$$\begin{bmatrix} \nabla^2 f(x_k) & J_{\mathcal{E}}(x_k)^T \\ J_{\mathcal{E}}(x_k) & 0 \end{bmatrix} \begin{bmatrix} p^\star \\ -y^\star \end{bmatrix} = - \begin{bmatrix} \nabla f(x_k) \\ c_{\mathcal{E}}(x_k) \end{bmatrix},$$

where $y^\star \in \mathbb{R}^{n_\mathcal{E}}$ is the vector of Lagrange multipliers. We can see that sequential quadratic programming requires to solve systems of the KKT form at each iteration. We refer the reader to Nocedal and Wright (2006) for more details.

### 1.2.4 Augmented Lagrangian method

We now consider another approach to solve general constrained optimization problems known as the *augmented Lagrangian method* (see Hestenes, 1969, and Powell, 1969). This technique consists in replacing a constrained optimization problem by an unconstrained one which combines the objective function and the constraint violation in some way. As in the previous section, we only focus on problems with equality constraints of the form (1.17) and we consider the approach introduced in this context by Nocedal and Wright (2006), Section 17.3.

Consider the Lagrangian function of problem (1.17),

$$\mathcal{L}(x, y) = f(x) - y^T c_{\mathcal{E}}(x), \tag{1.19}$$

where $y \in \mathbb{R}^{n_\mathcal{E}}$ is the vector of Lagrange multipliers. The method considers the Lagrangian function with a quadratic penalty term,

$$\begin{aligned} \mathcal{L}(x, y; \mu) &= \mathcal{L}(x, y) + \frac{\mu}{2} \|c_{\mathcal{E}}(x)\|_2^2 \\ &= f(x) - y^T c_{\mathcal{E}}(x) + \frac{\mu}{2} \|c_{\mathcal{E}}(x)\|_2^2, \end{aligned} \tag{1.20}$$

where $\mu > 0$ is called the *penalty parameter*. The first-order necessary optimality conditions given in Theorem 1.2, implies to solve the nonlinear equations $\nabla \mathcal{L}(x, y; \mu) = 0$, or, equivalently,

$$\begin{bmatrix} \nabla_x \mathcal{L}(x, y; \mu) \\ \nabla_y \mathcal{L}(x, y; \mu) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

As we have seen in Section 1.1.2, to find the roots of the nonlinear system $\nabla \mathcal{L}(x, y; \mu) = 0$, one can apply Newton's method and solve the Newton equations

$$\nabla^2 \mathcal{L}(x, y; \mu) p = -\nabla \mathcal{L}(x, y; \mu),$$

or, equivalently,

$$\begin{bmatrix} \nabla^2_{xx}\mathcal{L}(x,y;\mu) & \nabla^2_{xy}\mathcal{L}(x,y;\mu) \\ \nabla^2_{yx}\mathcal{L}(x,y;\mu) & \nabla^2_{yy}\mathcal{L}(x,y;\mu) \end{bmatrix} \begin{bmatrix} p_x \\ p_y \end{bmatrix} = - \begin{bmatrix} \nabla_x\mathcal{L}(x,y;\mu) \\ \nabla_y\mathcal{L}(x,y;\mu) \end{bmatrix}, \tag{1.21}$$

with $p_x \in \mathbb{R}^n$ and $p_y \in \mathbb{R}^{n_\mathcal{E}}$. The components of the gradient of the augmented Lagrangian (1.20) are given by

$$\begin{aligned} \nabla_x\mathcal{L}(x,y;\mu) &= \nabla f(x) - J_\mathcal{E}(x)^T y + \mu\, J_\mathcal{E}(x)^T c_\mathcal{E}(x) \\ &= \nabla_x\mathcal{L}(x,y) + \mu\, J_\mathcal{E}(x)^T c_\mathcal{E}(x), \end{aligned} \tag{1.22}$$

and

$$\nabla_y\mathcal{L}(x,y;\mu) = -c_\mathcal{E}(x), \tag{1.23}$$

while the components of its Hessian are given by

$$\nabla^2_{xx}\mathcal{L}(x,y;\mu) = \nabla^2_{xx}\mathcal{L}(x,y) + \mu\, J_\mathcal{E}(x)^T J_\mathcal{E}(x) + \mu \sum_{i\in\mathcal{E}} c_i(x)\nabla^2 c_i(x), \tag{1.24}$$

$$\nabla^2_{yy}\mathcal{L}(x,y;\mu) = 0_{n_\mathcal{E}}, \tag{1.25}$$

and

$$\nabla^2_{xy}\mathcal{L}(x,y;\mu) = -J_\mathcal{E}(x)^T. \tag{1.26}$$

Substituting in the Newton equations (1.21) yields the following KKT system

$$\begin{bmatrix} \nabla^2_{xx}\mathcal{L}(x,y;\mu) & -J_\mathcal{E}(x)^T \\ -J_\mathcal{E}(x) & 0_{n_\mathcal{E}} \end{bmatrix} \begin{bmatrix} p_x \\ p_y \end{bmatrix} = - \begin{bmatrix} \nabla_x\mathcal{L}(x,y) + \mu\, J_\mathcal{E}(x)^T c_\mathcal{E}(x) \\ -c_\mathcal{E}(x) \end{bmatrix},$$

where the $(1,1)$ block is given by (1.24).

As we have shown in this section and the previous one, we can solve optimization problems with equality constraints with two approaches: sequential quadratic programming or augmented Lagrangian[(1)]. In both cases, the key point is to solve systems of KKT form. In the next section, we consider an approach called interior-point method where we consider a constrained optimization problem with equality and inequality constraints.

[(1)] These methods can be adapted to problems with inequalities constraints. We refer the reader to the book of Nocedal and Wright (2006).

### 1.2.5 Interior-point method

In this section, we consider a constrained optimization problem with equality and inequality constraints,

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & c_{\mathcal{E}}(x) = 0 \\ & c_{\mathcal{I}}(x) \geq 0. \end{aligned} \tag{1.27}$$

We analyse a type of methods called *interior-point methods* suggested by Frisch (1955) and later developed by Fiacco and McCormick (1968), that require all iterates to strictly satisfy the inequality constraints in the problem and to respect the equality constraints. We refer the reader to the book on interior-point methods for linear programming (see, e.g., Wright, 1997) and to surveys on nonlinear optimization (see, e.g., Forsgren, Gill and Wright, 2002, Gould, Orban and Toint, 2005 and Nocedal and Wright, 2006). A subclass of interior-point methods adopting the most efficient practical approaches are known as primal-dual methods and were introduced in the early 1990s (see, e.g., Wright, 1997, and Nocedal and Wright, 2006).

As in Nocedal and Wright (2006), the inequality constraint of problem (1.27) can be reformulated as two constraints (one inequality and one equality) by adding a new variable $s \in \mathbb{R}^{n_{\mathcal{I}}}$ called a *slack variable*. One can indeed transform problem (1.27) into

$$\begin{aligned} \min_{x \in \mathbb{R}^n,\, s \in \mathbb{R}^{n_{\mathcal{I}}}} \quad & f(x) \\ \text{s.t.} \quad & c_{\mathcal{E}}(x) = 0 \\ & c_{\mathcal{I}}(x) - s = 0 \\ & s \geq 0. \end{aligned} \tag{1.28}$$

We then replace the non-negativity constraint in problem (1.28) by a logarithmic term called *barrier term* in the objective function,

$$\begin{aligned} \min_{x \in \mathbb{R}^n,\, s \in \mathbb{R}^{n_{\mathcal{I}}}} \quad & f(x) - \omega \sum_{i \in \mathcal{I}} \ln s_i \\ \text{s.t.} \quad & c_{\mathcal{E}}(x) = 0 \\ & c_{\mathcal{I}}(x) - s = 0, \end{aligned} \tag{1.29}$$

with $\omega > 0$ being a barrier parameter and $s_i$ being a component of $s$. The feasible set of problem (1.29) is the set of strictly feasible points for problem (1.28). The minimization of the barrier term $-\omega \sum_{i \in \mathcal{I}} \ln s_i$ in (1.29) prevents the components of $s$ from becoming too close to zero.

The Lagrangian function for problem (1.29) is given by

$$\mathcal{L}(x, s, y, z) = f(x) - \omega \sum_{i \in \mathcal{I}} \ln s_i - y^T c_{\mathcal{E}}(x) - z^T (c_{\mathcal{I}}(x) - s),$$

where $y \in \mathbb{R}^{n_\mathcal{E}}$ and $z \in \mathbb{R}^{n_\mathcal{I}}$ are the vectors of Lagrange multipliers. We write the KKT conditions (1.10)-(1.11) for problem (1.29) as follows,

$$\begin{aligned} \nabla_x \mathcal{L}(x, s, y, z) &= 0 \\ \nabla_s \mathcal{L}(x, s, y, z) &= 0 \\ c_{\mathcal{E}}(x) &= 0 \\ c_{\mathcal{I}}(x) - s &= 0, \end{aligned}$$

which is equivalent to $\nabla \mathcal{L}(x, s, y, z)$ where the gradient of the Lagrangian function is given by

$$\nabla \mathcal{L}(x, s, y, z) = \begin{bmatrix} \nabla_x \mathcal{L}(x, s, y, z) \\ \nabla_s \mathcal{L}(x, s, y, z) \\ \nabla_y \mathcal{L}(x, s, y, z) \\ \nabla_z \mathcal{L}(x, s, y, z) \end{bmatrix} = \begin{bmatrix} \nabla f(x) - J_{\mathcal{E}}(x)^T y - J_{\mathcal{I}}(x)^T z \\ -\omega S^{-1} e + z \\ -c_{\mathcal{E}}(x) \\ -c_{\mathcal{I}}(x) + s \end{bmatrix},$$

with $S$ being the diagonal matrix whose diagonal entries are given by the vector $s > 0$ while $e = [1, 1, \ldots, 1]^T \in \mathbb{R}^{n_\mathcal{I}}$. As we have seen in Section 1.1.2, to find the roots of the nonlinear system $\nabla \mathcal{L}(x, s, y, z) = 0$, one can solve the Newton equations. We first multiply the second equation $\nabla_s \mathcal{L}(x, s, y, z) = 0$ by $S$ implying

$$\nabla_s \mathcal{L}(x, s, y, z) = -\omega e + Sz = 0.$$

Applying Newton's method, we obtain

$$\nabla^2 \mathcal{L}(x, s, y, z) p = -\nabla \mathcal{L}(x, s, y, z)$$

or, equivalently, the next system to solve

$$\begin{bmatrix} \nabla^2_{xx} \mathcal{L}(x, s, y, z) & 0 & -J_{\mathcal{E}}(x)^T & -J_{\mathcal{I}}(x)^T \\ 0 & Z & 0 & S \\ -J_{\mathcal{E}}(x) & 0 & 0 & 0 \\ -J_{\mathcal{I}}(x) & I_{n_\mathcal{I}} & 0 & 0 \end{bmatrix} \begin{bmatrix} p_x \\ p_s \\ p_y \\ p_z \end{bmatrix} = - \begin{bmatrix} \nabla f(x) - J_{\mathcal{E}}(x)^T y - J_{\mathcal{I}}(x)^T z \\ Sz - \omega e \\ -c_{\mathcal{E}}(x) \\ -c_{\mathcal{I}}(x) + s \end{bmatrix},$$

with $Z$ being the diagonal matrix whose diagonal entries are given by the vector $z$, $p_x \in \mathbb{R}^n$, $p_s \in \mathbb{R}^{n_\mathcal{I}}$, $p_y \in \mathbb{R}^{n_\mathcal{E}}$ and $p_z \in \mathbb{R}^{n_\mathcal{I}}$. Multiplying the second equation by $S^{-1}$, we can rewrite the system in the symmetric form

$$\left[\begin{array}{ccc|cc} \nabla^2_{xx}\mathcal{L}(x,s,y,z) & 0 & & -J_{\mathcal{E}}(x)^T & -J_{\mathcal{I}}(x)^T \\ 0 & \Sigma & & 0 & I_{n_{\mathcal{I}}} \\ -J_{\mathcal{E}}(x) & 0 & & 0 & 0 \\ -J_{\mathcal{I}}(x) & I_{n_{\mathcal{I}}} & & 0 & 0 \end{array}\right] \begin{bmatrix} p_x \\ p_s \\ p_y \\ p_z \end{bmatrix} = - \begin{bmatrix} \nabla f(x) - J_{\mathcal{E}}(x)^T y - J_{\mathcal{I}}(x)^T z \\ z - \omega S^{-1} e \\ -c_{\mathcal{E}}(x) \\ -c_{\mathcal{I}}(x) + s \end{bmatrix}, \tag{1.30}$$

with the diagonal matrix $\Sigma = S^{-1}Z \in \mathbb{R}^{n_{\mathcal{I}} \times n_{\mathcal{I}}}$. Various formulations of the Newton equations can appear in the literature. For instance, the symmetric matrix (1.30) has a KKT form but we can consider two other formulations. First, in the second equation of (1.30), we extract

$$p_s = -\Sigma^{-1} p_z - \Sigma^{-1} z + \omega \Sigma^{-1} S^{-1} e,$$

and the system (1.30) can be reduced by eliminating $p_s$ with $Z^{-1} = \Sigma^{-1} S^{-1}$,

$$\left[\begin{array}{ccc} \nabla^2_{xx}\mathcal{L}(x,s,y,z) & -J_{\mathcal{E}}(x)^T & -J_{\mathcal{I}}(x)^T \\ -J_{\mathcal{E}}(x) & 0 & 0 \\ -J_{\mathcal{I}}(x) & 0 & -\Sigma^{-1} \end{array}\right] \begin{bmatrix} p_x \\ p_y \\ p_z \end{bmatrix} = - \begin{bmatrix} \nabla f(x) - J_{\mathcal{E}}(x)^T y - -J_{\mathcal{I}}(x)^T z \\ z - \omega S^{-1} e \\ -c_{\mathcal{I}}(x) + s - \Sigma^{-1} z + \omega Z^{-1} e \end{bmatrix}.$$

Observe that, by eliminating $p_z = -\Sigma J_{\mathcal{I}}(x) p_x - \Sigma c_{\mathcal{I}}(x) + \Sigma s - z + \omega \Sigma Z^{-1} e$ using the last equation, we reduce the matrix again to obtain the following KKT matrix,

$$\begin{bmatrix} \nabla^2_{xx}\mathcal{L}(x,s,y,z) + J_{\mathcal{I}}^T(x) \Sigma J_{\mathcal{I}}(x) & -J_{\mathcal{E}}(x)^T \\ -J_{\mathcal{E}}(x) & 0 \end{bmatrix}.$$

### 1.2.6 Regularization method

In this section, we analyse the cure of an ill-posed system. In some cases, the (1,1) block can be positive semidefinite and (1,2) block does not have a full column rank. This implies that the system is not invertible or that the solution of the system may not be unique. The *regularization* is a method that constructs a related problem whose solution is unique and only slightly differs from a solution of the original system. Regularization can take many forms and we analyse here the approach developed by Altman and Gondzio (1998) and Friedlander and Orban (2012)[(2)].

The first problem that can occur is that the matrix $A$ is not positive definite. In this case, we can replace the original matrix denoted $A$ by $A + \gamma I_n$ with the positive parameter $\gamma$. The modification of the (1,1) block implies to add $\gamma$ to the eigenvalues of $A$ and thus to induce the positive definiteness of the (1,1) block for $\gamma$ large enough. In the optimization context, it is similar to consider the *proximal-point method* developed by Rockafellar (1976) when the function $f$ is convex. This method generates a sequence of iterates that are approximate solutions by solving a sequence of subproblems of the form

[(2)]The authors use the regularization in the context of interior-point methods, while we focus only on problems with equality constraints.

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) + \frac{1}{2}\gamma_k \|x - x_k\|_2^2 \\ \text{s.t.} \quad & c_{\mathcal{E}}(x) = 0, \end{aligned} \tag{1.31}$$

where $\{\gamma_k\}$ is a sequence of decreasing positive parameters. The solution of problem (1.31) yields the next iterate $x_{k+1}$. A proximal-point term has been added to the original objective function, which penalizes solutions far from the previous iterate. The Lagrangian function of (1.31) is given by

$$\mathcal{L}(x, y) = f(x) + \frac{1}{2}\gamma_k \|x - x_k\|_2^2 - y^T c_{\mathcal{E}}(x),$$

where $y \in \mathbb{R}^{n_\mathcal{E}}$ is the vector of Lagrange multipliers. The KKT conditions (1.10)-(1.11) are given by

$$\begin{aligned} \nabla f(x) + \gamma_k(x - x_k) - J_{\mathcal{E}}(x)^T y &= 0 \\ c_{\mathcal{E}}(x) &= 0, \end{aligned}$$

and the Newton equations give

$$\begin{bmatrix} \nabla^2 f(x) + \gamma_k I_n & J_{\mathcal{E}}(x)^T \\ J_{\mathcal{E}}(x) & 0 \end{bmatrix} \begin{bmatrix} p_x \\ -p_y \end{bmatrix} = - \begin{bmatrix} \nabla f(x) + \gamma_k(x - x_k) - J_{\mathcal{E}}(x)^T y \\ c_{\mathcal{E}}(x) \end{bmatrix}. \tag{1.32}$$

The $(1,1)$ block of (1.32) is $\nabla^2 f(x) + \gamma_k I_n$, which is positive definite for $\gamma_k$ large enough, even if $\nabla^2 f(x)$ is positive semidefinite. It implies that the solution $x_{k+1}$ is unique when $J_{\mathcal{E}}(x)^T$ has full column rank. The conditions under which the proximal-point algorithm terminates under a finite number of iterations is given by Ferris (1991).

Afterwards, if the $(1,2)$ block has a deficient rank, the KKT matrix is singular. The solution for this problem is to modify the $(2,2)$ block by $-\delta z$ with the positive parameter $\delta$ called the *dual regularization parameter* and $z \in \mathbb{R}^{n_\mathcal{E}}$. Suppose that the objective function and the equality constraints of this problem are perturbed to yield the problem

$$\begin{aligned} \min_{x \in \mathbb{R}^n, z \in \mathbb{R}^{n_\mathcal{E}}} \quad & f(x) + \frac{1}{2}\delta \|z\|_2^2 \\ \text{s.t.} \quad & c_{\mathcal{E}}(x) + \delta z = 0. \end{aligned} \tag{1.33}$$

The Jacobian of the constraints (1.33) is $[J_{\mathcal{E}}(x) \quad \delta I_{n_\mathcal{E}}]$ that never has a deficient rank. The Lagrangian function of (1.33) is given by

$$\mathcal{L}(x,z,y) = f(x) + \frac{1}{2}\delta\|z\|_2^2 - y^T(c_{\mathcal{E}}(x) + \delta z),$$

where $y \in \mathbb{R}^{n_\varepsilon}$ is the vector of Lagrange multipliers. The KKT conditions (1.10)-(1.11) are given by

$$\begin{aligned} \nabla f(x) - J_{\mathcal{E}}(x)^T y &= 0 \\ \delta z - \delta y &= 0 \\ c_{\mathcal{E}}(x) + \delta z &= 0, \end{aligned}$$

implying the following Newton equations,

$$\left[\begin{array}{cc:c} \nabla^2_{xx}\mathcal{L}(x,z,y) & 0 & J_{\mathcal{E}}(x)^T \\ 0 & \delta I_{n_\varepsilon} & \delta I_{n_\varepsilon} \\ \hdashline J_{\mathcal{E}}(x) & \delta I_{n_\varepsilon} & 0 \end{array}\right] \begin{bmatrix} p_x \\ p_z \\ -p_y \end{bmatrix} = - \begin{bmatrix} \nabla f(x) - J_{\mathcal{E}}(x)^T y \\ \delta z - \delta y \\ c_{\mathcal{E}}(x) + \delta z \end{bmatrix},$$

with $p_x \in \mathbb{R}^n$, $p_z \in \mathbb{R}^{n_\varepsilon}$ and $p_y \in \mathbb{R}^{n_\varepsilon}$. By eliminating $p_z$ using the second equation, we obtain the following SQD matrix

$$\begin{bmatrix} \nabla^2_{xx}\mathcal{L}(x,z,y) & J_{\mathcal{E}}(x)^T \\ J_{\mathcal{E}}(x) & -\delta I_{n_\varepsilon} \end{bmatrix}.$$

The regularization implies to solve system of the KKT form or of the SQD form.

## 1.3 Properties of Karush-Kuhn-Tucker (KKT) matrices and of symmetric quasi-definite (SQD) matrices

In the previous section of this chapter, we saw that systems of the KKT or SQD form appear in various methods of optimization, such as sequential quadratic programming or interior-point methods. These systems form an important class of linear systems and it is crucial to solve them efficiently. In the next chapter, we analyse the techniques to solve this type of systems but before, we recall the properties of matrices of the KKT or SQD form as the invertibility conditions. We only focus on the solution of saddle-point systems (1.1) with $\mathcal{A}$ previously defined in (1.2) as

$$\mathcal{A} = \begin{bmatrix} A & B \\ B^T & -C \end{bmatrix}, \tag{1.34}$$

where $A$ of order $n$ is sparse, symmetric and positive definite, $B$ of size $n \times m$ has full column rank ($m \leq n$) and $C$ of order $m$ is either zero or symmetric and positive definite. The matrix $\mathcal{A}$ is then indefinite. More precisely, it has $n$ positive eigenvalues and $m$ negative eigenvalues, thanks to the following result.

**Theorem 1.6** If $A$ is positive definite and $C$ is positive semidefinite, then inertia($\mathcal{A}$)=$(n, m-p, p)$, where $0 \le p \le m$. If $B$ has full column rank or $C$ is positive definite then $p = 0$.

**Proof.** (Higham and Cheng, 1998, Lemma 4.2) □

The following result shows that the saddle-point matrix $\mathcal{A}$ with $A$ and $C$ positive semidefinite is nonsingular if the two row blocks $\begin{bmatrix} A & B^T \end{bmatrix}$ and $\begin{bmatrix} B & -C \end{bmatrix}$ have full row rank.

**Theorem 1.7** Assume that $A$ and $C$ are symmetric, positive semidefinite matrices. The vectors $u$ and $v$ satisfy

$$\begin{bmatrix} A & B \\ B^T & -C \end{bmatrix} \begin{bmatrix} u \\ -v \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

if and only if

$$\begin{bmatrix} A \\ B^T \end{bmatrix} u = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} B \\ -C \end{bmatrix} v = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

**Proof.** (Forsgren, Gill and Wong, 2015, Proposition 5) □

We consider the solution of KKT systems in (1.3),

$$\mathcal{A}_{KKT} x = b \equiv \begin{bmatrix} A & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}, \tag{1.35}$$

where $A \in \mathbb{R}^{n \times n}$ is symmetric and $B \in \mathbb{R}^{n \times m}$ $(m \le n)$. We first consider the case where the matrix $A$ is positive definite, for which the following result gives a necessary and sufficient condition for the saddle-point matrix $\mathcal{A}_{KKT}$ to be nonsingular.

**Theorem 1.8** Assume that $A$ is symmetric positive definite. Then the matrix $\mathcal{A}_{KKT}$ is nonsingular if and only $B$ has a full column rank.

**Proof.** See, e.g., (Benzi et al., 2005, Theorem 3.1) □

In the particular case where $A$ is positive semidefinite, we have the following result.

**Theorem 1.9** Assume that $A$ is symmetric positive semidefinite and $B$ has full column rank. Then the KKT matrix $\mathcal{A}_{KKT}$ is nonsingular if and only if

$$Ker(A) \cap Ker(B^T) = \{0\}.$$

**Proof.** See, e.g., (Benzi et al., 2005, Theorem 3.2) □

The requirement that $A$ be positive semidefinite can be somewhat relaxed by the next theorem.

**Theorem 1.10** Assume that $A$ is positive definite on $Ker(B^T)$ and $B$ has a full column rank. Then the KKT matrix $\mathcal{A}_{KKT}$ is nonsingular.

**Proof.** See, e.g., (Nocedal and Wright, 2006, Lemma 16.1) □

In this work, we assume that $A$ is symmetric positive definite and $B$ has a full column rank for the KKT matrices. By Theorem 1.8, the matrix $A_{KKT}$ is nonsingular and thus the systems (1.35) has a unique solution.

Finally, we consider a general SQD system in (1.4),

$$\mathcal{A}_{SQD}x = b \equiv \begin{bmatrix} A & B \\ B^T & -C \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix},$$

where $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{n\times n}$ are symmetric positive definite and $B \in \mathbb{R}^{n\times m}$ $(m \leq n)$. The SQD matrices $\mathcal{A}_{SQD}$ are always nonsingular such that the inverse of SQD matrices has a particular form.

**Theorem 1.11** The inverse of a SQD matrix is SQD.

**Proof.** (Vanderbei, 1995, Theorem 1) □

We recall an important result, which shows that the symmetric quasi-definite matrices form a class of strongly factorizable matrices where a permutation matrix is defined by permuting the rows or columns of an $n \times n$ identity matrix according to some permutation of the numbers 1 to $n$.

**Theorem 1.12** The SQD matrices are strongly factorizable, i.e., there exists a factorization $P\mathcal{A}_{SQD}P^T = LDL^T$ for any permutation $P$ with $D$ a diagonal matrix and $L$ a unit lower triangular matrix.

**Proof.** (Vanderbei, 1995, Theorem 2) □

Chapter 2

# Introduction to iterative methods for solving systems of equations

As we have seen in the first chapter, the solution of linear systems of equations is of great interest in many numerical optimization approaches. In the first part of this chapter, we consider a general linear system of equations

$$Ax = b, \tag{2.1}$$

where the square matrix $A$ of order $n$ is symmetric. When $A$ has a small size, direct methods based on Gaussian elimination are generally used. Such methods involve a fixed number of steps that require a finite number of operations and at the end, provide the solution. For symmetric positive definite matrices, for instance one can use the *Cholesky factorization*, which requires $\frac{n^3}{3}$ operations in exact arithmetic and transforms $A$ into the product of a lower triangular matrix $L$ and its conjugate transpose. The factors $L$ and $L^T$ are used to solve (2.1) through the solution of two triangular systems easier to solve: $Ly = b$ by forward substitution, followed by the solution of $Ux = y$ via backward substitution. If the matrix coefficient is sparse, some fill-in can appear during the factorization and, in addition, when $A$ is large, direct methods are too costly, which motivates the use of iterative techniques to find a good approximation of the solution. We refer the reader to the books of Greenbaum (1997) and Hageman and Young (1981) for more details on iterative methods for solving linear systems.

We introduce here the Krylov subspace methods which aim to solve linear systems (2.1), in particular, three iterative methods which belong to the class of the Krylov methods: the *Lanczos method*, the *conjugate gradient method* and the *minimal residuals method*. The last two are respectively known as the CG method and the MINRES method. The CG method is the most popular

method used for symmetric positive definite systems of the form (2.1), while the MINRES method is developed for symmetric indefinite ones.

In the second part of this chapter, we present the general preconditioning techniques for linear systems, which aim to accelerate the convergence of iterative methods. Especially, we focus on the preconditioners for the KKT systems or the SQD systems as for instance the block diagonal preconditioners or the constraint preconditioners. We also analyse in more details the block diagonal preconditioner proposed by Golub et al. (2006).

## 2.1 Krylov subspace methods

The main idea of *Krylov subspace methods* is to consider an initial iterate $x_0 \in \mathbb{R}^n$ with the *initial residual* $r_0 := Ax_0 - b$ and to generate a sequence of iterates such that the $k$th iterate $x_k$ satisfies

$$x_k \in x_0 + \mathcal{K}(A, x_0; k),$$

where

$$\mathcal{K}(A, x; k) := span\{x, A\,x, \ldots, A^{k-1}x\},$$

is the *Krylov subspace* of degree $k \leq n$ for $x$. The dimension of these subspaces increases by one at each iteration of the method. The construction of the iterates is based on an orthonormal basis $\{v_1, \ldots, v_k\}$ of the Krylov subspace so that the approximate solution at the $k$th iteration is given by

$$x_k = x_0 + V_k y_k, \tag{2.2}$$

where $y_k \in \mathbb{R}^k$ and $V_k \in \mathbb{R}^{n\times k}$ is the matrix with columns $v_1, \ldots, v_k$. Krylov methods can roughly be classified in four families depending on the manner in which they compute $x_k$. Literature on Krylov subspace methods can be found in van der Vorst (2003) and Saad (2003). Before presenting some of them, we recall for further use how an orthonormal basis of the Krylov subspace can be built.

The *Lanczos method* introduced by Lanczos (1952), has initially been developed to compute a few dominant eigenvalues and possibly the associated eigenvectors of a large sparse symmetric matrix $A$. However it also builds an orthonormal basis of a Krylov subspace $\mathcal{K}(A, x; k)$ whose basis vectors can be expressed in terms of polynomials in the matrix $A$ applied to the initial vector $x_0$ (see, e.g., Meurant and Strakoš, 2006).

Let $v_1$ be an initial vector and let $v_0 = 0$. Based on the Gram-Schmidt algorithm, an orthonormal basis $\{v_1, v_2, \ldots, v_k\}$ for $\mathcal{K}(A, v_1; k)$ can be constructed by the following algorithm (see, e.g., Fischer, 2011, p.135).

**Algorithm 1** Lanczos method

1: Choose an initial vector $v_1$;
2: Set $\gamma_1 = \|v_1\|_2$, $v_0 = 0$;
3: **for** $j = 1, 2, \ldots$ **do**
4: $\quad v_j = v_j/\gamma_j$;
5: $\quad \delta_j = v_j^T A v_j$;
6: $\quad v_{j+1} = Av_j - \delta_j v_j - \gamma_j v_{j-1}$;
7: $\quad \gamma_{j+1} = \|v_{j+1}\|_2$;
8: **end for**

The basis vectors $\{v_i\}_{i=1}^k$ are known as *Lanczos vectors* and the Lanczos method transforms a symmetric matrix $A$ into a symmetric tridiagonal matrix $T_{k+1,k}$ with an additional row at the bottom,

$$T_{k+1,k} = \begin{bmatrix} \delta_1 & \gamma_2 & 0 & \cdots & 0 \\ \gamma_2 & \delta_2 & \gamma_3 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \gamma_k \\ 0 & \cdots & 0 & \gamma_k & \delta_k \\ 0 & \cdots & \cdots & 0 & \gamma_{k+1} \end{bmatrix} \in \mathbb{R}^{(k+1)\times k}.$$

If we define $T_k$ to be the first $k$ rows of $T_{k+1,k}$, then $T_k$ is square, symmetric and

$$T_{k+1,k} = \begin{bmatrix} T_k \\ \gamma_{k+1} e_k^T \end{bmatrix},$$

where $e_k$ is the $k$th column of the $k \times k$ identity matrix $I_k$. Observe that line 6 of Algorithm 1 for $j = 1, \ldots, k$ can be written in matrix form as

$$\begin{aligned} AV_k &= V_{k+1} T_{k+1,k} \\ &= V_k T_k + \gamma_{k+1} v_{k+1} e_k^T. \end{aligned} \tag{2.3}$$

From the orthogonality of the Lanczos vectors, we have that $V_k^T v_{k+1} = 0$ and we can deduce the following relation

$$V_k^T A V_k = T_k, \tag{2.4}$$

which will be useful in the next sections. In the next sections, we introduce the general principle of three methods belonging to the class of Krylov subspace methods.

### 2.1.1 Lanczos algorithm

When solving linear systems (2.1) with $A$ symmetric, the approximation of the solution is given by (2.2). To identify suitable iterates $x_k$, we first consider the *Ritz-Galerkin approach*, also used by the CG method, which imposes that the *Ritz-Galerkin condition*, i.e., that the *residual* $r_k := Ax_k - b$ be orthogonal to the subspace $\mathcal{K}(A, r_k; k) = span\{r_k, A\,r_k, \ldots, A^{k-1}r_k\}$. Replacing the expression (2.2) in the $k$th residual gives

$$r_k = A(x_0 + V_k y_k) - b,$$

or, equivalently,

$$r_k = r_0 + AV_k y_k. \tag{2.5}$$

Substituting the previous expression in the Ritz-Galerkin condition $V_k^T r_k = 0$ together with the relation (2.4) implies that

$$V_k^T r_0 + V_k^T A V_k y_k = 0$$

and we deduce that

$$y_k = -T_k^{-1}(V_k^T r_0).$$

with $T_k$ nonsingular. We set $v_1 = r_0/\beta$ with $\beta = \|r_0\|_2$ and we obtain

$$y_k = -T_k^{-1}(\beta e_1), \tag{2.6}$$

where $e_1$ is the first column of the $k \times k$ identity matrix $I_k$. The pseudocode for the Lanczos method to solve linear systems is given by Algorithm 2. Note that lines 1-9 correspond to the Lanczos method (Algoritm 1) applied with $v_1 = r_0$ and line 10 corresponds to the approximation of the solution given by (2.2) with (2.6).

**Algorithm 2** Lanczos method for linear systems

1: Choose an initial vector $x_0$;
2: Compute $r_0 = Ax_0 - b$, $v_1 = r_0$;
3: Set $\gamma_1 = \|v_1\|_2$, $v_0 = 0$;
4: **for** $j = 1, 2, \ldots$ **do**
5: $\quad v_j = v_j/\gamma_j$;
6: $\quad \delta_j = v_j^T A v_j$;
7: $\quad v_{j+1} = Av_j - \delta_j v_j - \gamma_j v_{j-1}$;
8: $\quad \gamma_{j+1} = \|v_{j+1}\|_2$;
9: **end for**
10: Compute $y_k = -T_k^{-1}(\gamma_1 e_1)$ and set $x_k = x_0 + V_k y_k$;

### 2.1.2 Conjugate gradient (CG) algorithm

The CG algorithm introduced by Hestenes and Stiefel (1952), is an iterative method to solve linear systems (2.1) where the matrix $A$ is symmetric and positive definite. Although the CG algorithm was first derived in a completely different way using conjugacy and minimization of a quadratic function, it turns out that it is mathematically equivalent to the Lanczos algorithm described in the previous section (see, e.g., Meurant, 2006, Section 2.1). Indeed, it can be obtained from the Lanczos algorithm by using the Cholesky factorization of $T_k$.

We introduce here the CG algorithm as a method which minimizes the quadratic function $q(x) := \frac{1}{2}x^T Ax - b^T x$, where $A$ is symmetric and positive definite. The minimizer of this function is the solution to $Ax = b$, as we have seen in Section 1.1.1. We initialize for a given starting point $x_0 \in \mathbb{R}^n$, the initial residual $r_0 := Ax_0 - b$ and the descent direction $p_0 = -r_0$. The CG method generates the sequence of iterates $\{x_k\}$ by setting

$$x_{k+1} = x_k + \alpha_k p_k,$$

where $p_k$ is a descent direction at iteration $k$ and $\alpha_k$ is the step length determined by an exact linesearch along $p_k$ (i.e., the minimizer of $q(.)$ along $p_k$). It is given explicitly by

$$\alpha_k = -\frac{r_k^T p_k}{p_k^T A p_k}. \tag{2.7}$$

The new search direction $p_k$ is generated using only the previous direction $p_{k-1}$ and is defined as a linear combination of the residual $r_k$ and $p_{k-1}$,

$$p_k = -r_k + \beta_k p_{k-1}, \tag{2.8}$$

where the scalar $\beta_k$ is chosen to ensure that $p_{k-1}$ and $p_k$ are $A$-conjugate, i.e.,

$$p_{k-1}^T A p_k = 0.$$

We set

$$\beta_k = \frac{r_k^T A p_{k-1}}{p_{k-1}^T A p_{k-1}}. \tag{2.9}$$

Doing so, one can show that all the generated directions are $A$-conjugate,

$$p_i^T A p_j = 0 \quad \text{for all} \quad i \neq j.$$

The pseudocode for the CG method is presented in Algorithm 3 (see, e.g., Nocedal and Wright, 2006, p.112). Line 4 and line 7 compute the step length and the scalar $\beta_k$ at iteration $k$ based on and equivalent to formula (2.7) and (2.9), respectively.

**Algorithm 3** CG method

Choose an initial vector $x_0$;
Set $r_0 = Ax_0 - b$, $p_0 = -r_0$, $k = 0$;
**while** $r_k \neq 0$ **do**
    $\alpha_k = \dfrac{r_k^T r_k}{p_k^T A p_k}$;
    $x_{k+1} = x_k + \alpha_k p_k$;
    $r_{k+1} = r_k + \alpha_k A p_k$;
    $\beta_{k+1} = \dfrac{r_{k+1}^T r_{k+1}}{r_k^T r_k}$;
    $p_{k+1} = -r_{k+1} + \beta_{k+1} p_k$;
    $k = k + 1$;
**end while**

In exact arithmetic, the CG method converges in at most $n$ iterations where $n$ is the order of the matrix $A$, but in finite precision, it could not be the case. Indeed, the performance of the CG method depends on the distribution of the eigenvalues and/or on the *condition number* of the matrix $A$ defined, for a nonsingular matrix, as $\kappa(A) = \|A\|\|A^{-1}\|$, where any matrix norm can be used. In the case of the Euclidiean norm, we have for a symmetric positive definite matrix that $\kappa_2(A) = \frac{\lambda_n}{\lambda_1}$, with $\lambda_1$ and $\lambda_n$ the smallest and the largest eigenvalues of $A$, respectively.

The following theorem says that the more eigenvalues are clustered, the more convergence is rapid.

**Theorem 2.1** If $A$ has only $r$ distinct eigenvalues, then in exact arithmetic, the CG method will terminate at the solution in at most $r$ iterations.

**Proof.** See, e.g., (Nocedal and Wright, 2006, p.115) □

The two following theorems give a bound on the error norm of iterate with respect to the eigenvalues or the condition number of $A$. The *A-norm* of the vector $x \in \mathbb{R}^n$ is defined by $\|x\|_A = \sqrt{x^T A x}$.

**Theorem 2.2** If $A$ has eigenvalues $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$, we have that

$$||x_{k+1} - x^*||_A^2 \leq \left(\frac{\lambda_{n-k} - \lambda_1}{\lambda_{n-k} + \lambda_1}\right)^2 ||x_0 - x^*||_A^2. \tag{2.10}$$

**Proof.** (Luenberger, 1973, p.180) □

The $A$-norm of the error of iterate can also be bounded by the following convergence theorem.

**Theorem 2.3**

$$||x_{k+1} - x^*||_A \leq 2 \left( \frac{\sqrt{\kappa_2(A)} - 1}{\sqrt{\kappa_2(A)} + 1} \right)^{k+1} ||x_0 - x^*||_A. \qquad (2.11)$$

**Proof.** See, e.g., (Conn et al., 2000, p.85) □

The convergence of the CG method is not only affected by the condition number of $A$ but also by the number and distribution of very small eigenvalues, as shown by van der Sluis and van der Vorst (1986).

### 2.1.3 Minres algorithm

The Minres algorithm was derived by Paige and Saunders (1975) and can be viewed as a generalization of the CG method for the solution of symmetric indefinite linear systems. One way to derive the Minres algorithm is to exploit the minimum norm residual approach which implies that the Euclidean norm of the residual $\|r_k\|_2$, is minimal over the Krylov subspace $\mathcal{K}(A, v_1; k)$ with $v_1 = r_0/\beta$ and $\beta = \|r_0\|_2$. Similarly to Section 2.1.1, the $k$th residual (2.5) is given by

$$r_k = r_0 + AV_k y_k,$$

where the columns of $V_k$ are the Lanczos vectors generated by the Lanczos algorithm (Algorithm 1) and $y_k \in \mathbb{R}^k$ such that

$$y_k = \arg \min_{y \in \mathbb{R}^k} \|r_0 + AV_k y\|_2.$$

Using expression (2.3), we obtain

$$\begin{aligned} r_k &= r_0 + V_{k+1} T_{k+1,k} y_k \\ &= V_{k+1}(\beta e_1 + T_{k+1,k} y_k). \end{aligned}$$

with $e_1$ is the first column of the $(k+1) \times (k+1)$ identity matrix $I_{k+1}$. Since the columns of the matrix $V_{k+1}$ are orthonormal, we have

$$\|r_k\|_2 = \|\beta e_1 + T_{k+1,k} y_k\|_2,$$

so that the approximate solution at the $k$th iteration is given by

$$x_k = x_0 + V_k y_k, \tag{2.12}$$

where $y_k = \arg\min_{y \in \mathbb{R}^k} \|\beta e_1 + T_{k+1,k} y\|_2$, which is the solution of a least-squares problem.

The usual technique to solve this least-squares problem is to use a $QR$ factorization which transforms the tridiagonal matrix $T_{k+1,k}$ into the product of an orthogonal matrix $Q \in \mathbb{R}^{(k+1)\times(k+1)}$ and an upper bidiagonal form[(1)] $R \in \mathbb{R}^{(k+1)\times k}$ by using Givens rotations (see line 18 in Algorithm 4 below). We define $(QR)_{k,k}$ to be the first $k$ rows and columns of $QR$. The solution $y_k$ is obtained by solving the upper bidiagonal system

$$\beta e_1 + (QR)_{k,k}\, y_k = 0,$$

which implies that

$$y_k = -\beta R_{k,k}^{-1} (Q^T e_1)_{1:k}, \tag{2.13}$$

where $R_{k,k}$ is the upper $k$-by-$k$ block of the tridiagonal factor $R$ and $(Q^T e_1)_{1:k}$ are the $k$ first entries of vector $Q^T e_1 \in \mathbb{R}^{k+1}$. Using the expression (2.13) in the approximate solution (2.12), we obtain

$$x_k = x_0 - \beta\, (V_k R_{k,k}^{-1})\, (Q^T e_1)_{1:k}.$$

In practice, we set $W_k := V_k R_{k,k}^{-1}$ and the columns of $W_k$ denoted by $\{w_j\}_{j=1}^k$ are computed in Algorithm 4 by line 20 below. The pseudocode of the MINRES method is described in Algorithm 4 and we refer the reader to Saad (2003) and Greenbaum (1997) for more details.

We have a result for the 2-norm of the residual in the MINRES algorithm. As for the CG algorithm, the bound depends on the condition number of the matrix $A$.

**Theorem 2.4**

$$||Ax_k - b||_2 \leq 2 \left( \frac{\sqrt{\kappa_2(A)} - 1}{\sqrt{\kappa_2(A)} + 1} \right)^k ||Ax_0 - b||_2. \tag{2.14}$$

(1) That is, the top $k$-by-$k$ block is upper bidiagonal and the last row is zero.

**Algorithm 4** Minres method

Initialization
$v_0 = 0$, $w_0 = 0$, $w_1 = 0$
Choose an initial vector $x_0$, compute $v_1 = b - Ax_0$
set $\gamma_1 = \sqrt{v_1^T v_1}$
Set $\eta = \gamma_1$, $s_0 = s_1 = 0$, $c_0 = c_1 = 1$
**for** $j = 1$ until convergence **do**
  Lanczos
    $v_j = v_j/\gamma_j$
    $\delta_j = v_j^T A v_j$
    $v_{j+1} = Av_j - \delta_j v_j - \gamma_j v_{j-1}$
    $\gamma_{j+1} = \|v_{j+1}\|_2$
  QR factorization
    $\alpha_0 = c_j\delta_j - c_{j-1}s_j\gamma_j$
    $\alpha_1 = \sqrt{\alpha_0^2 + \gamma_{j+1}^2}$
    $\alpha_2 = s_j\delta_j + c_{j-1}c_j\gamma_j$
    $\alpha_3 = s_{j-1}\gamma_j$
  Givens rotation
    $c_{j+1} = \alpha_0/\alpha_1$; $s_{j+1} = \gamma_{j+1}/\alpha_1$
  Least squares solution
    $w_{j+1} = (v_j - \alpha_3 w_{j-1} - \alpha_2 w_j)/\alpha_1$
    $x_j = x_{j-1} + c_{j+1}\eta w_{j+1}$
    $\eta = -s_{j+1}\eta$ with $|\eta| = \|Ax_j - b\|$
**end for**

**Proof.** See, e.g., (Greenbaum, 1997, p.53) □

### 2.1.4 Comparison between CG and Minres

In the previous sections, we have presented two famous Krylov subspace methods, CG and Minres, which are applied on positive definite or indefinite symmetric systems, respectively. Indeed, the CG method can be unstable or undefined on systems that are not positive definite. The natural choice for saddle-point systems of KKT or SQD form is thus the Minres method. The bounds (2.10) and (2.11) for CG method and (2.14) for Minres method, lead to the observation that if a matrix $A$ has a small condition number and/or that its eigenvalues are clustered, then the convergence of the CG or Minres method will be rapid.

One way to classify and compare these Krylov subspace methods is based on the quantity to be minimized. Table 2.1 summarizes the quantity to be minimized for each method with the error norm defined by $e_k := x_k - x^\star$.

| | $k$th residual | $k$th error |
|---|---|---|
| CG with $A$ positive definite | min $\|r_k\|_{A^{-1}}$ | min $\|e_k\|_A$ |
| MINRES | min $\|r_k\|_2$ | |

Table 2.1 – Residual and error properties of CG and MINRES

The efficiency of iterative techniques can be improved by using preconditioning, which is simply a mean of transforming the original linear system into another one having the same solution but which is likely to be more rapidly solved with an iterative solver. The next section introduces the general technique of preconditioning.

## 2.2 Preconditioning

The important feature of the CG and MINRES methods is that at each iteration, only one matrix times vector multiplication and a small number of vector operations are required. For sparse or structured matrices, the matrix times vector product may be efficiently computed. In iterative methods, the total computational work to solve a linear system hence essentially depends on the number of iterations it takes to have convergence with an acceptable accuracy.

Preconditioning is usually crucial to ensure that this number is kept acceptably small. A *preconditioner* transforms the linear system into another equivalent one that has better spectral properties, as these impact the convergence rate. Typically, large spreads and little clustering in the spectrum of $A$ lead to a slow convergence of the iterative methods. In practice, a good preconditioner should be cheap to construct and to apply. It corresponds to the application of a non singular matrix $P \in \mathbb{R}^{n\times n}$ to the original linear system to yield a different linear system for which the convergence of the iterative method will be significantly faster. One can think of preconditioned iteration as applying the original iteration to the system

$$P^{-1}Ax = P^{-1}b. \tag{2.15}$$

However the application of $P$ as in (2.15) would be a bad choice since we would then create a non-symmetric linear system whereas $A$ is originally symmetric.

In general, the iterative solution of non-symmetric linear systems is more expensive and one tries to preserve symmetry. If $P$ is symmetric and positive definite, we can write $P = LL^T$ for some matrix $L$ (e.g., either the Cholesky factor or the matrix square root). The iterative method is then applied to the symmetric system

$$L^{-1}AL^{-T}y = L^{-1}b \qquad \text{where} \qquad L^Tx = y$$

and convergence depends on the eigenvalues of the symmetric and positive definite matrix $L^{-1}AL^{-T}$. Benzi (2002) gives a nice survey on preconditioning

techniques for large linear systems. The pseudocode of the *preconditioned conjugate gradient method* (PCG) is given by Algorithm 5 (see, e.g., Nocedal and Wright, 2006, p.119).

**Algorithm 5** Preconditioned conjugate gradient method

Choose an initial vector $x_0$;
Choose a preconditioner symmetric positive definite $P$;
Set $r_0 = Ax_0 - b$;
Solve $Py_0 = r_0$ for $y_0$;
Set $p_0 = -y_0$, $k = 0$;
**while** $r_k \neq 0$ **do**
    $\alpha_k = \frac{r_k^T y_k}{p_k^T A p_k}$;
    $x_{k+1} = x_k + \alpha_k p_k$;
    $r_{k+1} = r_k + \alpha_k A p_k$;
    Solve $Py_{k+1} = r_{k+1}$;
    $\beta_{k+1} = \frac{r_{k+1}^T y_{k+1}}{r_k^T y_k}$;
    $p_{k+1} = -y_{k+1} + \beta_{k+1} p_k$;
    $k = k + 1$;
**end while**

For Minres, a symmetric and positive definite preconditioner $P$ must also be employed and the convergence will depend on the eigenvalues of the symmetric and indefinite matrix $L^{-1}AL^{-T}$. The pseudocode of the preconditioned Minres method is described in Algorithm 6 (see, e.g., Elman et al., 2005, Section 6.1). In the next sections, we introduce possible approaches to precondition indefinite symmetric systems of KKT or SQD form.

In the context of very large saddle-point systems, preconditioning techniques for Krylov subspace methods are very useful. The paper of Benzi and Wathen (2008) gives an overview of the most useful preconditioning techniques for Krylov subspace methods applied to saddle-point problems, including block-diagonal preconditioners and constraint preconditioners. Before introducing these two types of preconditioners, we recall the spectral properties of the KKT and SQD matrices and then, for the following of the thesis, we denote the matrices of KKT or SQD form by

$$\mathcal{A}_{KKT} := \begin{bmatrix} A & B \\ B^T & 0 \end{bmatrix} \text{ and } \mathcal{A}_{SQD} := \begin{bmatrix} A & B \\ B^T & -C \end{bmatrix}, \tag{2.16}$$

where $A$ of order $n$ is symmetric and positive definite, $B$ of size $n \times m$ has a full column rank ($m \leq n$) and $C$ of order $m$ is symmetric and positive definite. As we have seen in the previous sections, the convergence of some iterative methods mainly depends on the eigenvalues distribution of the system. The spectrum of the KKT matrix $\mathcal{A}_{KKT}$ contains both positive and negative eigenvalues as shown by Theorem 1.6, and Rusten and Winther (1992) have established an

**Algorithm 6** Preconditioned MINRES method

INITIALIZATION
$v_0 = 0$, $w_0 = 0$, $w_1 = 0$
Choose an initial vector $x_0$;
Choose a preconditioner symmetric positive definite $P$;
Ccompute $v_1 = b - Ax_0$
Solve $Pz_1 = v_1$, set $\gamma_1 = \sqrt{v_1^T z_1}$
Set $\eta = \gamma_1$, $s_0 = s_1 = 0$, $c_0 = c_1 = 1$
**for** $j = 1$ until convergence **do**
  LANCZOS
    $z_j = z_j/\gamma_j$
    $\delta_j =< Az_j, z_j >$
    $v_{j+1} = Az_j - \delta_j v_j - \gamma_j v_{j-1}$
    solve $Pz_{j+1} = v_{j+1}$
    $\gamma_{j+1} = \sqrt{v_{j+1}^T z_{j+1}}$
  QR FACTORIZATION
    $\alpha_0 = c_j\delta_j - c_{j-1}s_j\gamma_j$
    $\alpha_1 = \sqrt{\alpha_0^2 + \gamma_{j+1}^2}$
    $\alpha_2 = s_j\delta_j + c_{j-1}c_j\gamma_j$
    $\alpha_3 = s_{j-1}\gamma_j$
  GIVENS ROTATION
    $c_{j+1} = \alpha_0/\alpha_1$; $s_{j+1} = \gamma_{j+1}/\alpha_1$
  UPDATE
    $w_{j+1} = (z_j - \alpha_3 w_{j-1} - \alpha_2 w_j)/\alpha_1$
    $x_j = x_{j-1} + c_{j+1}\eta w_{j+1}$
    $\eta = -s_{j+1}\eta$
**end for**

important result relating the spectrum of $\mathcal{A}_{KKT}$ to the eigenvalues of $A$ and to the singular values of $B$.

**Theorem 2.5** Assume $A$ is symmetric positive definite and $B$ has a full column rank. Let $0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ be the eigenvalues of $A$ and $0 < \sigma_1 \leq \sigma_2 \leq \ldots \leq \sigma_m$ be the singular values of $B$. Then the eigenvalues of $\mathcal{A}_{KKT}$ are bounded within

$$\left[\frac{\lambda_1 - \sqrt{\lambda_1^2 + 4\sigma_m^2}}{2}, \frac{\lambda_n - \sqrt{\lambda_n^2 + 4\sigma_1^2}}{2}\right] \cup \left[\lambda_1, \frac{\lambda_n + \sqrt{\lambda_n^2 + 4\sigma_m^2}}{2}\right].$$

**Proof.** (Rusten and Winther, 1992, Lemma 2.1) □

This theorem has been extended to the case of the SQD system by Silvester and Wathen (1994).

**Theorem 2.6** Assume $A$ is symmetric positive definite, $B$ has a full column rank and $C$ is symmetric semi-positive definite. Let $0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ be the eigenvalues of $A$, $0 < \sigma_1 \leq \sigma_2 \leq \ldots \leq \sigma_m$ be the singular values of $B$ and $0 \leq \lambda_1^C \leq \lambda_2^C \leq \ldots \leq \lambda_m^C$ be the eigenvalues of $C$. Then the eigenvalues of $\mathcal{A}_{SQD}$ are bounded within

$$\left[\frac{\lambda_1 - \lambda_m^C - \sqrt{(\lambda_1 + \lambda_m^C)^2 + 4\sigma_m^2}}{2}, \frac{\lambda_n - \sqrt{\lambda_n^2 + 4\sigma_1^2}}{2}\right] \cup \left[\lambda_1, \frac{\lambda_n + \sqrt{\lambda_n^2 + 4\sigma_m^2}}{2}\right].$$

**Proof.** (Silvester and Wathen, 1994, Lemma 2.2) □

The only difference between the bounds in Theorems 2.5 and 2.6 is in the lower bound on the negative eigenvalues (left interval). We can see that the largest eigenvalue of $C$ appears twice in the numerator.

In Section 2.2.1, we analyse eigenvalues distribution of the KKT or SQD matrices preconditioned by the block diagonal preconditioners. We recall in Section 2.2.2, the preconditioner introduced by Golub et al. (2006) and in Section 2.2.3, we give our contribution to this preconditioner. Finally, we introduce constraint preconditioners in Section 2.2.4.

### 2.2.1 Block diagonal preconditioners

If $A$ is nonsingular, we can decompose the saddle-point matrix $\mathcal{A}$ into the following block triangular factorization,

$$\begin{bmatrix} A & B \\ B^T & -C \end{bmatrix} = \begin{bmatrix} I_n & 0 \\ B^T A^{-1} & I_m \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & -(C + B^T A^{-1} B) \end{bmatrix} \begin{bmatrix} I_n & A^{-1}B \\ 0 & I_m \end{bmatrix}, \quad (2.17)$$

where $C + B^T A^{-1} B$ is called the *Schur complement*[(2)] of the saddle-point matrix and is denoted by $S$. If we assume that $A$ and $C$ are positive definite, then the Schur complement $C + B^T A^{-1} B$ is also positive definite. Since the matrix $\mathcal{A}$ has a block structure, it makes sense for the preconditioner $\mathcal{P}$ to also have a block structure. The approach that we propose in this thesis is based

[(2)] We draw attention on the fact that we use the term "Schur complement" corresponding to the opposite of the formal definition of the Schur complement defined by $-C - B^T A^{-1} B$. The goal of which is to simplify and improve the clarity of this work.

on a well-known result of Murphy et al. (2000). The authors consider, when $C = 0$, the "ideal" block diagonal preconditioner,

$$\mathcal{P} := \begin{bmatrix} A & 0 \\ 0 & S \end{bmatrix}, \tag{2.18}$$

in which the exact Schur complement $S := B^T A^{-1} B$ is used. The following result gives the eigenvalues distribution of the preconditioned system $\mathcal{P}^{-1}\mathcal{A}_{KKT}$ of the form

$$\mathcal{P}^{-1}\mathcal{A}_{KKT} = \begin{bmatrix} I_n & A^{-1}B \\ S^{-1}B^T & 0 \end{bmatrix}.$$

**Theorem 2.7** Let $\mathcal{P}$ be given by (2.18) with the Schur complement $S := B^T A^{-1} B$. Then the preconditioned matrix $\mathcal{P}^{-1}\mathcal{A}_{KKT}$ has exactly three distinct eigenvalues,

$$1, \frac{1+\sqrt{5}}{2}, \frac{1-\sqrt{5}}{2}. \tag{2.19}$$

**Proof.** See Murphy et al. (2000), Proposition 1 when the preconditioned matrix $\mathcal{P}^{-1}\mathcal{A}_{KKT}$ is nonsingular. □

A similar result may be obtained for $C \neq 0$ and positive semidefinite, given by Gould and Simoncini (2009). In this case, the "ideal" block diagonal preconditioner (2.18) in which the exact Schur complement $S := B^T A^{-1} B + C$ is used, yields a preconditioned matrix $\mathcal{P}^{-1}\mathcal{A}_{SQD}$ of the form

$$\mathcal{P}^{-1}\mathcal{A}_{SQD} = \begin{bmatrix} I_n & A^{-1}B \\ S^{-1}B^T & -S^{-1}C \end{bmatrix}.$$

**Theorem 2.8** Let $\mathcal{P}$ be given by (2.18) with the Schur complement $S := B^T A^{-1} B + C$. Then the eigenvalues of the preconditioned matrix $\mathcal{P}^{-1}\mathcal{A}_{SQD}$ are equal to (2.19) or

$$\frac{1}{2\theta}\left(\theta - 1 \pm \sqrt{(1-\theta)^2 + 4\theta^2}\right),$$

where $\theta$ denotes the eigenvalues of the generalized eigenvalue problem

$$\left(B^T A^{-1} B + C\right) u \;=\; \theta\, Cu.$$

**Proof.** (Gould and Simoncini, 2009, Proposition 4.2) □

In practice, these ideal preconditioners are too expensive to compute and to apply and we use approximations of both blocks $A$ and $S$ to compute an approximation denoted by

$$\tilde{\mathcal{P}} := \begin{bmatrix} \tilde{A} & 0 \\ 0 & \tilde{S} \end{bmatrix}. \tag{2.20}$$

In Chapter 3, we introduce a new approach to approximate the two blocks $A$ and $S$ using spectral information extracted from the matrix $A$. For the following of this chapter, we only consider the KKT systems. We introduce the block diagonal preconditioner given by Golub et al. (2006) leading to our theoretical contribution.

### 2.2.2 Golub-Greif-Varah ($\mathcal{GGV}$) preconditioner

In Golub et al. (2006), the authors assume that $A$ is only positive semidefinite and replace the $(1,1)$ block $A$ of the KKT system by $A + BWB^T$ with $W \in \mathbb{R}^{m\times m}$, a symmetric positive semidefinite matrix. The system (1.1) with the matrix coefficient $\mathcal{A}_{KKT}$ in (2.16) is thus transformed into the following system

$$\begin{bmatrix} A + BWB^T & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} f + BWB^T u \\ g \end{bmatrix}$$

or, equivalently,

$$\begin{bmatrix} A + BWB^T & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} f + BWg \\ g \end{bmatrix}. \tag{2.21}$$

The authors consider the block diagonal preconditioner as the "ideal" preconditioner defined by

$$\begin{bmatrix} A + BWB^T & 0 \\ 0 & B^T(A + BWB^T)^{-1}B \end{bmatrix} \tag{2.22}$$

associated to the system (2.21), where $B^T(A + BWB^T)^{-1}B$ is the Schur complement. The benefit of such an approach is that the $(1,1)$ block of the modified linear system (2.21) may be made nonsingular, hence positive definite, and well-conditioned. Note that, in Golub et al. (2006), the choice $W = \omega I_m$ is considered, with $\omega$ a positive scalar such that the linear system becomes

$$\begin{bmatrix} A + \omega BB^T & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} f + \omega Bg \\ g \end{bmatrix}, \tag{2.23}$$

and the "ideal" associated preconditioner ($\mathcal{G}$olub, $\mathcal{G}$reif, $\mathcal{V}$arah) becomes

$$\mathcal{P}_{\mathcal{GGV}} = \begin{bmatrix} A + \omega BB^T & 0 \\ 0 & B^T(A + \omega BB^T)^{-1}B \end{bmatrix}. \tag{2.24}$$

The authors perform an algebraic study of such a preconditioning approach, showing how the eigenvalues of the preconditioned matrix $\mathcal{P}_{\mathcal{GGV}}{}^{-1}\mathcal{A}_{KKT}$ are clustered in some interval whose ends are isolated from the origin and well bounded towards infinity. They also construct approximations of the block diagonal preconditioner (2.24) by explicitly building some approximations of the Schur complement $B^T \left(A + \omega BB^T\right)^{-1} B$ in which $A + \omega BB^T$ is replaced, for instance, by its diagonal part or its incomplete Cholesky decomposition (see, e.g., Greenbaum, 1997, Section 11.1). In practice, they illustrate the convergence curves of MINRES for the solution of saddle-point systems with different approximations of the preconditioner (2.24).

As mentioned at the end of Section 3 in Golub et al. (2006), the authors set

$$\omega = \frac{\|A\|_2}{\|B\|_2^2}, \tag{2.25}$$

but without motivating this choice. In the framework of our theoretical and empirical study on block diagonal preconditioners, we wished to motivate this choice of $\omega$. We provide, in the next section, a theoretical result that we have derived about the value of $\omega$.

### 2.2.3 A theoretical contribution to the $\mathcal{GGV}$ preconditioner

As far as we know, no theoretical result exists in the literature to justify the choice (2.25) for the value of the parameter $\omega$ in (2.24). We establish below a new result given one approach to show that the choice of $\omega$ can be motivated by the minimization of the condition number of the matrix $A + \omega BB^T$.

**Theorem 2.9** Let $A+\omega BB^T \in \mathbb{R}^{n\times n}$ with $A$ symmetric positive semidefinite and singular, $B \in \mathbb{R}^{n\times m}$ has full column rank ($m \leq n$) with $0 < \sigma_1 \leq \sigma_2 \leq \ldots \leq \sigma_m$ be the singular values of $B$ and $w \geq 0$. Let $\xi_1 \in \mathbb{R}^m$. Then

$$\kappa_2(A + \omega BB^T) \geq h(\omega), \tag{2.26}$$

with $h(\omega) = \max(f_1(\omega), f_2(\omega), f_4(\omega))$ where

$$\begin{aligned} f_1(\omega) &= \frac{\lambda_{\max}(A) + \omega\|\xi_1\|_2^2\sigma_1^2}{\omega\sigma_m^2}, \\ f_2(\omega) &= \frac{\|\xi_1\|_2^2\sigma_1^2}{\sigma_m^2}, \\ f_4(\omega) &= \frac{\lambda_{\max}(A) + \omega\|\xi_1\|_2^2\sigma_1^2}{\lambda_{\max}(A)}, \end{aligned}$$

and we have that

$$\omega^* := \arg\min_{\omega\in\mathbb{R}^+} h(\omega) = \frac{\|A\|_2}{\|B\|_2^2}.$$

**Proof.** The proof is divided in three parts. The first one uses the singular value decomposition of $B$ to imply the existence of an orthogonal matrix $P \in \mathbb{R}^{n\times n}$, to transform the matrix $P^T\left(A + \omega BB^T\right)P$, which is similar to $A+\omega BB^T$, into a matrix with block structure. In the second part of the proof, we analyse the Rayleigh quotient of $P^T\left(A + \omega BB^T\right)P$ to deduce some bounds on the extreme eigenvalues of this matrix. Finally, the last part studies these bounds to deduce the desired result.

The first part of the proof relies on the *singular value decomposition* (SVD) of $B$ (see Appendix A) that guarantees the existence of orthogonal matrices $P \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{m\times m}$ such that

$$P^T BV = \Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_m) \in \mathbb{R}^{n\times m} \qquad (m \leq n) \tag{2.27}$$

where $\{\sigma_i\}_{i=1}^m$ are the singular values of $B$ satisfying $0 < \sigma_1 \leq \sigma_2 \leq \ldots \leq \sigma_m$. Isolating $B$ from (2.27) and considering the following decomposition of $P$,

$$P = \left[\begin{array}{cc} P_B & P_{\bar{B}} \end{array}\right] \tag{2.28}$$

with the columns of the matrix $P_B \in \mathbb{R}^{n\times m}$ in $\mathcal{I}m(B)$ and the columns of the matrix $P_{\bar{B}} \in \mathbb{R}^{n\times(n-m)}$ in $(\mathcal{I}m(B))^\perp = \mathcal{K}er(B^T)$, lead to the thin SVD of $B$ (see, e.g., Golub and Van Loan, 2013, Section 2.4.3),

$$\begin{aligned} B &= P\Sigma V^T, \\ &= \left[\begin{array}{cc} P_B & P_{\bar{B}} \end{array}\right] \left[\begin{array}{c} \Sigma_1 \\ 0_{(n-m)\times m} \end{array}\right] V^T, \\ &= P_B \Sigma_1 V^T, \end{aligned} \tag{2.29}$$

with $\Sigma_1 = \mathrm{diag}(\sigma_1, \ldots, \sigma_m) \in \mathbb{R}^{m\times m}$. Substituting now (2.28) for $P$ in the matrix $P^T\left(A + \omega BB^T\right)P$ (which is similar to $A + \omega BB^T$) and using the fact that the columns of $P_{\bar{B}} \in \mathcal{K}er(B^T)$, we obtain

$$\begin{aligned} P^T\left(A + \omega BB^T\right)P &= \left[\begin{array}{c} P_B^T \\ P_{\bar{B}}^T \end{array}\right] \left(A + \omega BB^T\right) \left[\begin{array}{cc} P_B & P_{\bar{B}} \end{array}\right], \\ &= \left[\begin{array}{cc} P_B^T(A + \omega BB^T)P_B & P_B^T A P_{\bar{B}} \\ P_{\bar{B}}^T A P_B & P_{\bar{B}}^T A P_{\bar{B}} \end{array}\right]. \end{aligned}$$

Using (2.29), we observe that $BB^T = (P_B\Sigma_1 V^T)(P_B\Sigma_1 V^T)^T = P_B\Sigma_1^2 P_B^T$, implying that

$$\left[\begin{array}{cc} P_B^T(A + \omega BB^T)P_B & P_B^T A P_{\bar{B}} \\ P_{\bar{B}}^T A P_B & P_{\bar{B}}^T A P_{\bar{B}} \end{array}\right] = \left[\begin{array}{cc} P_B^T A P_B + \omega\Sigma_1^2 & P_B^T A P_{\bar{B}} \\ P_{\bar{B}}^T A P_B & P_{\bar{B}}^T A P_{\bar{B}} \end{array}\right],$$

by orthonormality of the columns of $P_B$.

For the second part of the proof, consider, for any non-zero vector $\xi \in \mathbb{R}^n$ having the partition $\left[\begin{array}{c} \xi_1 \\ \xi_2 \end{array}\right]$ with $\xi_1 \in \mathbb{R}^m$ and $\xi_2 \in \mathbb{R}^{(n-m)}$, the Rayleigh quotient

$$r(\xi) = \frac{\left[\begin{array}{cc} \xi_1^T & \xi_2^T \end{array}\right] \left[\begin{array}{cc} P_B^T A P_B + \omega\Sigma_1^2 & P_B^T A P_{\bar{B}} \\ P_{\bar{B}}^T A P_B & P_{\bar{B}}^T A P_{\bar{B}} \end{array}\right] \left[\begin{array}{c} \xi_1 \\ \xi_2 \end{array}\right]}{\left[\begin{array}{cc} \xi_1^T & \xi_2^T \end{array}\right] \left[\begin{array}{c} \xi_1 \\ \xi_2 \end{array}\right]}. \tag{2.30}$$

Without loss of generality, we consider that the vector $\xi$ is unit ($\|\xi\|_2 = 1$) and we denote by $\lambda_{\min}$ and $\lambda_{\max}$ the smallest and the largest eigenvalues of $P^T(A + \omega BB^T)P$ (and hence of $A + \omega BB^T$), respectively, implying that

$$\lambda_{\min} \leq r(\xi) \leq \lambda_{\max}, \tag{2.31}$$

by the Rayleigh-Ritz theorem (see, e.g., Horn and Johnson, 1985, Section 4.2). Using (2.30), we can rewrite (2.31) as

$$\lambda_{\min} \leq \left[\begin{array}{cc} \xi_1^T & \xi_2^T \end{array}\right] \left[\begin{array}{cc} P_B^T A P_B & P_B^T A P_{\bar{B}} \\ P_{\bar{B}}^T A P_B & P_{\bar{B}}^T A P_{\bar{B}} \end{array}\right] \left[\begin{array}{c} \xi_1 \\ \xi_2 \end{array}\right] + \omega\xi_1^T\Sigma_1^2\xi_1 \leq \lambda_{\max}, \tag{2.32}$$

or, equivalently,

$$\lambda_{\min} \leq \xi_1^T P_B^T A P_B \xi_1 + \xi_1^T P_B^T A P_{\bar{B}} \xi_2 \quad + \quad \xi_2^T P_{\bar{B}}^T A P_B \xi_1 \quad + \quad \xi_2^T P_{\bar{B}}^T A P_{\bar{B}} \xi_2 + \omega \xi_1^T \Sigma_1^2 \xi_1 \leq \lambda_{\max}. \tag{2.33}$$

As (2.33) is valid for all non-zero $\xi \in \mathbb{R}^n$, we will select four specific configurations which lead to some bounds on $\lambda_{\min}$ and $\lambda_{\max}$. These bounds will be exploited in the last part to conclude the proof.

a) In the first case, we consider $\xi_1 \in \mathbb{R}^m$ and $\xi_2 = 0$. The relation (2.33) becomes

$$\lambda_{\min} \leq \xi_1^T P_B^T A P_B \xi_1 + \omega \xi_1^T \Sigma_1^2 \xi_1 \leq \lambda_{\max}.$$

Since the matrix $A$ is symmetric positive semidefinite and $P_B$ has a full column rank, we get that $P_B^T A P_B$ is symmetric positive semidefinite and we thus have

$$\omega \xi_1^T \Sigma_1^2 \xi_1 \leq \xi_1^T P_B^T A P_B \xi_1 + \omega \xi_1^T \Sigma_1^2 \xi_1 \leq \lambda_{\max}.$$

Since $w \xi_1^T \Sigma_1^2 \xi_1 \geq \omega \sigma_1^2 \|\xi_1\|_2^2$, we obtain the following inequality

$$\omega \|\xi_1\|_2^2 \sigma_1^2 \leq \lambda_{\max}. \tag{2.34}$$

b) For the second case, we consider $\xi_1 = 0$ and $\xi_2 \in \mathbb{R}^{n-m}$. The relation (2.33) becomes

$$\lambda_{\min} \leq \xi_2^T P_{\bar{B}}^T A P_{\bar{B}} \xi_2 \leq \lambda_{\max}.$$

Since $\xi_2^T P_{\bar{B}}^T A P_{\bar{B}} \xi_2 \leq \lambda_{\max}(P_{\bar{B}}^T A P_{\bar{B}}) \|\xi_2\|_2^2$ and $\|\xi_2\|_2 \leq 1$ ($\xi$ being a unit vector), we obtain the following inequality

$$\lambda_{\min} \leq \xi_2^T P_{\bar{B}}^T A P_{\bar{B}} \xi_2 \leq \lambda_{\max}(P_{\bar{B}}^T A P_{\bar{B}}),$$

and thus

$$\lambda_{\min} \quad \leq \quad \lambda_{\max}(P_{\bar{B}}^T A P_{\bar{B}}),$$

where $\lambda_{\max}(P_{\bar{B}}^T A P_{\bar{B}})$ denotes the largest eigenvalue of $P_{\bar{B}}^T A P_{\bar{B}}$. We obtain the next inequality by the fact that the supremum of a function over $\mathbb{R}^n$ is greater than or equal to the supremum over the set of vectors of the form $x = P_{\bar{B}} y$, $y \in \mathbb{R}^{n-m}$,

$$\lambda_{\max}(P_{\bar{B}}^T A P_{\bar{B}}) := \max_{y \in \mathbb{R}^{n-m}} \frac{y^T P_{\bar{B}}^T A P_{\bar{B}} y}{y^T y} \quad \leq \quad \lambda_{\max}(A) := \max_{x \in \mathbb{R}^n} \frac{x^T A x}{x^T x},$$

by orthonormality of the columns of $P_{\bar{B}}$. We thus have

$$\lambda_{\min} \leq \lambda_{\max}(A). \tag{2.35}$$

c) Now, let us consider for the third case that $\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = P^T \begin{bmatrix} \bar{\xi}_1 \\ \bar{\xi}_2 \end{bmatrix}$, where $\bar{\xi}$ is the unit eigenvector associated to the eigenvalue of $A$ equal to zero. The relation (2.32) becomes

$$\lambda_{\min} \quad \leq \quad 0 + \omega \xi_1^T \Sigma_1^2 \xi_1 \leq \lambda_{\max}.$$

Since $w \xi_1^T \Sigma_1^2 \xi_1 \leq \omega \sigma_m^2 \|\xi_1\|_2$ with $\|\xi_1\|_2 \leq 1$ ($\xi$ being a unit vector), we obtain the following inequality

$$\lambda_{\min} \leq \omega \sigma_m^2. \tag{2.36}$$

d) The last case we consider, is the vector $\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = P^T \begin{bmatrix} \tilde{\xi}_1 \\ \tilde{\xi}_2 \end{bmatrix}$, where $\tilde{\xi}$ corresponds to the unit eigenvector associated to the largest eigenvalue of $A$. The inequality (2.32) then becomes

$$\lambda_{\min} \leq \lambda_{\max}(A) + \omega \xi_1^T \Sigma_1^2 \xi_1 \leq \lambda_{\max}.$$

Since $\omega \xi_1^T \Sigma_1^2 \xi_1 \geq \omega \sigma_1^2 \|\xi_1\|_2^2$, we have

$$\lambda_{\max}(A) + \omega \|\xi_1\|_2^2 \sigma_1^2 \leq \lambda_{\max}. \tag{2.37}$$

In last part of the proof, we finally derive some lower bounds on the condition number of $A + \omega B B^T$ by combining the previous inequalities (2.34), (2.35), (2.36) and (2.37). Respectively, by (2.36) and (2.37), by (2.34) and (2.36), by (2.34) and (2.35), and by (2.35) and (2.37), we obtain

$$\kappa_2(A + \omega B B^T) \geq \max\left(f_1(\omega), f_2(\omega), f_3(\omega), f_4(\omega)\right),$$

where

$$\begin{aligned} f_1(\omega) &= \frac{\lambda_{\max}(A) + \omega\|\xi_1\|_2^2\sigma_1^2}{\omega\sigma_m^2}, \\ f_2(\omega) &= \frac{\|\xi_1\|_2^2\sigma_1^2}{\sigma_m^2}, \\ f_3(\omega) &= \frac{\omega\|\xi_1\|_2^2\sigma_1^2}{\lambda_{\max}(A)}, \\ f_4(\omega) &= \frac{\lambda_{\max}(A) + \omega\|\xi_1\|_2^2\sigma_1^2}{\lambda_{\max}(A)}. \end{aligned} \tag{2.38}$$

We can notice that $f_3(\omega)$ and $f_4(\omega)$ are linear functions with a slope equal to $\frac{\|\xi_1\|_2^2\sigma_1^2}{\lambda_{\max}(A)}$ and $f_2(\omega)$ is a constant function. The function $f_3(\omega)$ is always below the function $f_4(\omega)$, implying the following bound on the condition number,

$$\kappa_2(A + \omega BB^T) \geq \max\left(f_1(\omega), f_2(\omega), f_4(\omega)\right). \tag{2.39}$$

We now analyse the two possible configurations of functions $f_1(\omega)$, $f_2(\omega)$ and $f_4(\omega)$ defined in (2.38). We consider the configuration where the intersections between $f_1(\omega)$ and $f_4(\omega)$ is below $f_2(\omega)$ in the left-hand subplot in Figure 2.1 or these intersection is above $f_2(\omega)$ in the right-hand subplot. The functions $f_1(\omega)$, $f_2(\omega)$, $f_3(\omega)$ and $f_4(\omega)$ are plotted in blue, green, red and black respectively for $\omega \in [0, 1]$ with $f_1(\omega) = 1/\omega$, $f_2(\omega) = 3$, $f_3(\omega) = 4\ \omega$ and $f_4(\omega) = 1 + 4\ \omega$ for the first configuration and we change $f_3(\omega) = 25\ \omega$ and $f_4(\omega) = 1 + 25\ \omega$ for the second one. The grey area in Figure 2.1 represents the set of values of $\kappa_2(A + \omega BB^T)$ satisfying the bounds (2.39) in both configurations. The three intersection points denoted by $I_1$, $I_2$ and $I_3$ and belonging to the grey area, are potential candidats to the problem

$$\min_{\omega \in \mathbb{R}^+} h(w),$$

with $h(\omega) := \max\left(f_1(\omega), f_2(\omega), f_4(\omega)\right)$.
Observe that the intersection between $f_1(\omega)$ and $f_2(\omega)$, denoted by $I_1$, implies that

$$\lambda_{\max}(A) = 0. \tag{2.40}$$

Hence we have that $A$ is a null matrix, which is impossible. On the other hand, the intersection between $f_2(\omega)$ and $f_4(\omega)$, denoted by $I_2$, implies that

$$\frac{\|\xi_1\|_2^2\sigma_1^2}{\sigma_m^2} = \frac{\lambda_{\max}(A) + \omega\|\xi_1\|_2^2\sigma_1^2}{\lambda_{\max}(A)},$$

or equivalently,

$$\omega = \frac{\lambda_{\max}(A)\|\xi_1\|_2^2\sigma_1^2 - \sigma_m^2\lambda_{\max}(A)}{\sigma_1^2\sigma_m^2\|\xi_1\|_2^2},$$

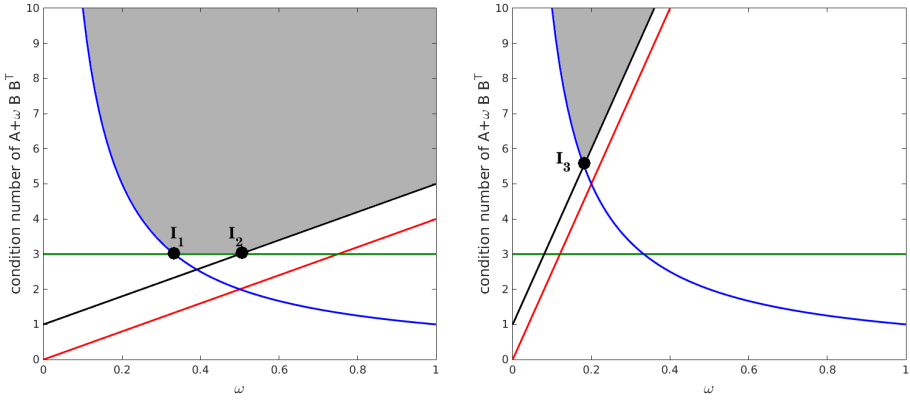



Figure 2.1 – Bounds on $\kappa_2(A + \omega BB^T)$.

which leads to

$$\omega = \frac{\lambda_{\max}(A)}{\sigma_m^2} \left( \frac{\|\xi_1\|_2^2 \sigma_1^2 - \sigma_m^2}{\|\xi_1\|_2^2 \sigma_1^2} \right). \tag{2.41}$$

Since $\|\xi_1\|_2^2 \leq 1$ and $\sigma_1^2 \leq \sigma_m^2$, we have $\|\xi_1\|_2^2 \sigma_1^2 - \sigma_m^2 \leq 0$, which is impossible since $\omega \geq 0$. The last possibility is the intersection between $f_1(\omega)$ and $f_4(\omega)$, denoted by $I_3$ that implies that

$$\frac{\lambda_{\max}(A) + \omega \|\xi_1\|_2^2 \sigma_1^2}{\omega \sigma_m^2} = \frac{\lambda_{\max}(A) + \omega \|\xi_1\|_2^2 \sigma_1^2}{\lambda_{\max}(A)},$$

which by simplification yields as only positive solution

$$\omega = \frac{\lambda_{\max}(A)}{\sigma_m^2}. \tag{2.42}$$

Since $A$ is symmetric, the 2-norm of $A$ is equal to the largest eigenvalue of $A$ (see, e.g., Golub and Van Loan, 2013, Section 2.3.3) and the 2-norm of $B$ is equal to the largest singular value of $B$ (see, e.g., Golub and Van Loan, 2013, Section 2.4.2), proving the result. □

We now illustrate the effect of different values of $\omega$ on the condition number of the saddle-point system genhs28 from the CUTEr test set Gould, Orban and Toint (2001*b*), as in Golub et al. (2006). This matrix is a $18 \times 18$ saddle-point matrix where $A$ is $10 \times 10$ and $B$ is $10 \times 8$. Figure 2.2 shows the evolution of the condition number of the matrix obtained by adding a multiple of $BB^T$ to $A$. We indicate the value of $\omega = \|A\|_2 / \|B\|_2^2 = 0.2308$ by a red star. We can

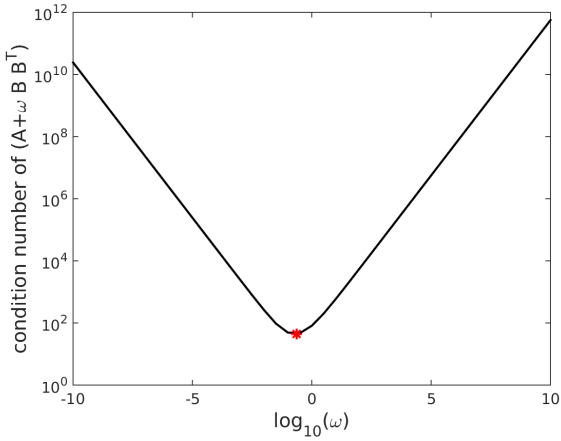



Figure 2.2 – Condition number of $A + \omega BB^T$ for the genhs28 matrix from CUTEr.

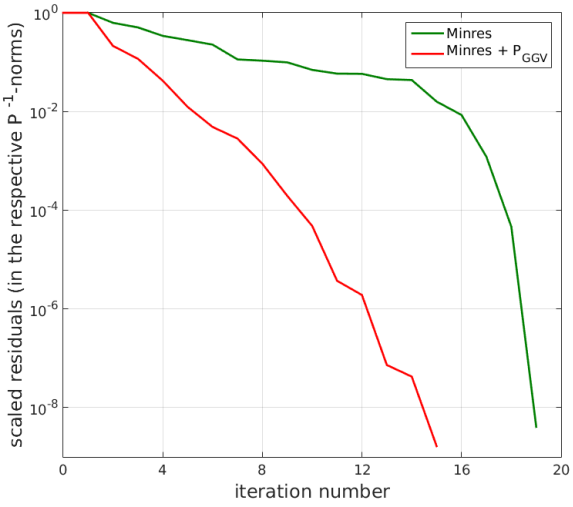



Figure 2.3 – Convergence profiles of Minres for the genhs28 matrix from CUTEr.

see that the condition number starts to decrease towards the condition number associated to $\omega = \|A\|_2/\|B\|_2^2$ before increasing again.
Figure 2.3 shows the convergence profile of preconditioned Minres with the preconditioner (2.24) using $\omega = 0.2308$ and Minres without preconditioner for the above test problem. The iterations are stopped when the scaled residuals

in respectively 2-norm and $\mathcal{P}^{-1}$ norm, are less than $10^{-8}$ and we can observe that the convergence with preconditioning is faster.

In the case where $A$ has high nullity, Greif and Schötzau (2006) replace the Schur complement in (2.22) by the symmetric positive definite matrix $W$ introduced in Section 2.2.2. They show that the preconditioned saddle-point system has $n$ eigenvalues equal to 1 and $m$ equal to $-1$, which implies that the preconditioned MINRES is expected to converge within two iterations.

Finally, in Golub and Greif (2003), the authors seek for a value of $\omega$ large enough so as to eliminate the effect of the ill-conditioning of $A$, while not too large to avoid the effect of the ill-conditioning of $BB^T$. They analyse the condition number of the saddle-point system (2.23) and show that the condition number gets larger when $\omega$ increases, and behaves like $\omega^2$.

### 2.2.4 Constraint preconditioners

In this last section, we consider another type of preconditioners for solving linear systems of the KKT form. This is a nonsingular preconditioner, called *constraint preconditioner* of the form

$$\mathcal{P}_c = \begin{bmatrix} G & B \\ B^T & 0 \end{bmatrix}, \tag{2.43}$$

where $G \in \mathbb{R}^{n\times n}$ is an approximation of the matrix $A$. The constraint preconditioner for indefinite linear systems was studied by Keller, Gould and Wathen (2000), or see Benzi and Wathen (2008) for a survey. We introduce it in this work for comparison purposes with the block diagonal preconditioners that we develop in Chapter 4.

Note that the blocks of preconditioner $\mathcal{P}_c$ in (2.43) are unchanged from the original matrix $\mathcal{A}_{KKT}$ in (2.16) so that the preconditioner $\mathcal{P}_c$ is an indefinite matrix as is $\mathcal{A}_{KKT}$. This implies that the MINRES method is not appropriate in this case. The authors in Gould, Hribar and Toint (2001*a*) show how the CG method combined only with the constraint preconditioner can still be used on indefinite linear systems of the KKT form. This is a real advantage for this preconditioner since the CG method is a very efficient method.

A first observation is that we need to solve only one system with the matrix $\mathcal{P}_c$ in (2.43) at each iteration of the CG algorithm. For instance, Dollar and Wathen (2004) use a new factorization for the preconditioned step of CG based on Schilders' factorization. In exact arithmetic, a second advantage of this preconditioner is that the CG combined with the constraint preconditioner ensures that all the iterates satisfy the constraints and this is not the case for other preconditoners.

Keller et al. (2000) give the next result on the eigenvalues distribution of the preconditioned matrix $\mathcal{P}_c^{-1}\mathcal{A}_{KKT}$. We include it here for completeness.

**Theorem 2.10** Let $\mathcal{A}_{KKT} \in \mathbb{R}^{(n+m)\times(n+m)}$ be the symmetric and indefinite matrix defined in (2.16). Assume $Z \in \mathbb{R}^{n\times(n-m)}$ is a basis for the null-space of $B^T$. Preconditioning $\mathcal{A}_{KKT}$ by the constraint preconditioner $\mathcal{P}_c$ defined in (2.43) where $G \neq A$, implies that the matrix $\mathcal{P}_c^{-1}\mathcal{A}_{KKT}$ has

1. an eigenvalue at 1 with multiplicity $2m$, and
2. $n - m$ eigenvalues which are defined by the generalized eigenvalue problem $Z^T AZx = \lambda Z^T GZx$.

**Proof.** (Keller et al., 2000, Theorem 2.1) □

We have introduced in this chapter the iterative methods to linear systems. In the case where we consider the KKT systems or the SQD systems, we have analysed in particular, the block diagonal preconditioner built on the matrix $A$ and on the Schur complement. In the next chapter, we introduce the inverse approximations of $A$ and of the Schur complement.

Chapter 3

# Spectral preconditioners for positive definite matrices

As we have seen in Chapter 2, preconditioning techniques are used to accelerate Krylov subspace methods and in particular, we have introduced in Section 2.2.1 the "ideal" block diagonal preconditioner,

$$\mathcal{P} := \begin{bmatrix} A & 0 \\ 0 & S \end{bmatrix}, \tag{3.1}$$

where $S = B^T A^{-1} B + C$ is the exact Schur complement. Following the ideas in Murphy et al. (2000), our goal in this chapter is to introduce good approximations of the inverse of the $(1,1)$ block $A$ and of the Schur complement $S$ used in Chapter 4 to build some appropriate preconditioners of the form (3.1) for a KKT or a SQD system with matrix, respectively,

$$\mathcal{A}_{KKT} := \begin{bmatrix} A & B \\ B^T & 0 \end{bmatrix} \text{ and } \mathcal{A}_{SQD} := \begin{bmatrix} A & B \\ B^T & -C \end{bmatrix}, \tag{3.2}$$

where $A$ of order $n$ is symmetric and positive definite, $B$ of size $n \times m$ has a full column rank and $C$ of order $m$ is symmetric and positive definite. Indeed, the starting point of this work was actually the former work that appeared in Giraud et al. (2006) and Golub et al. (2007), giving the ground basis for extracting and exploiting spectral information in the context of the solution of linear symmetric positive definite systems. We analyse thereafter how this could be extended to KKT or SQD systems, which exhibit very particular algebraic structures that can be exploited in a specific manner.

We assume that $A$ is ill-conditioned and that some first level of preconditioning has been applied to the systems (3.2) so that the spectrum of $A$ is clustered, with relatively few very small eigenvalues. This situation occurs when considering usual preconditioning techniques on $A$ such as, for instance,

the incomplete Cholesky decomposition (see, e.g., Greenbaum, 1997, Section 11.1) or a Jacobi scaling. For simplicity, we shall use $\mathcal{A}_{KKT}$ or $\mathcal{A}_{SQD}$ with matrices $A$, $B$ and $C$ as defined in (3.2) to refer to the KKT matrix or SQD matrix with the first level of preconditioning in use.

In a first stage, we also initially assume that we know the few very small eigenvalues and associated eigenvectors of $A$. Based on this knowledge, we thus aim at performing a further level of preconditioning on the system (3.2) that ensures, when the first level is not satisfying, a sufficiently fast convergence of MINRES. We shall derive our preconditioners from prior spectral information extracted from $A$ directly, more precisely from the subspace associated with the smallest eigenvalues of $A$. One of the benefits of our approach is that it allows us to work separately on $A$ and $B$, recombining them through the Schur complement approximation. This aspect will be studied in more details in Chapter 5.

The chapter is organized as follows. In Sections 3.1 and 3.2, we introduce the inverse approximations of $A$ and of the Schur complement using spectral information we will consider, and we study the spectral properties of the preconditioned matrices using these approximations. These spectral approximations will be used in Chapter 4 to build two block diagonal preconditioners for the KKT matrix or SQD matrix.

## 3.1 Spectral approximation of the inverse of the (1,1) block

Let the eigendecomposition of the matrix $A$ in (3.2) be given by

$$A = U\Lambda U^T = U_\gamma \Lambda_\gamma U_\gamma^T + \tilde{U}_\gamma \tilde{\Lambda}_\gamma \tilde{U}_\gamma^T, \tag{3.3}$$

where the spectrum $\{\lambda_i\}_{i=1}^n$ of $A$ is split in two parts, with $\Lambda_\gamma \in \mathbb{R}^{p\times p}$ the diagonal matrix containing the $p$ eigenvalues less than a given positive number $\gamma \in [\lambda_{\min}(A), \lambda_{\max}(A)]$, and with $\tilde{\Lambda}_\gamma \in \mathbb{R}^{(n-p)\times(n-p)}$ the diagonal matrix containing all the other $(n-p)$ eigenvalues. The columns of the rectangular matrices $U_\gamma \in \mathbb{R}^{n\times p}$ and $\tilde{U}_\gamma \in \mathbb{R}^{n\times(n-p)}$ are the orthonormal sets of eigenvectors corresponding to $\Lambda_\gamma$ and $\tilde{\Lambda}_\gamma$ respectively and form the orthogonal matrix $U = [U_\gamma,\ \tilde{U}_\gamma] \in \mathbb{R}^{n\times n}$. We assume that $U_\gamma$ and $\Lambda_\gamma$ are available[(1)].

Let $\alpha > 0$ be some known estimate of the average of the eigenvalues in $\tilde{\Lambda}_\gamma$ ( or of $\lambda_{\max}(A)$). Consider now the approximate inverse of $A$ given by the spectral low rank update (SLRU) approach developed by Carpentieri et al. (2003),

$$A_\gamma^{-1} = U_\gamma \Lambda_\gamma^{-1} U_\gamma^T + \frac{1}{\alpha} I_n\,. \tag{3.4}$$

[(1)] or that some good approximations can be computed, for instance by the approach proposed by Golub et al. (2007) and described in Chapter 7.

The eigenvalues $\{\mu_i\}_{i=1}^n$ of the matrix $A_\gamma^{-1}A$, with

$$\begin{aligned} A_\gamma^{-1}A &= \left(U_\gamma \Lambda_\gamma^{-1} U_\gamma^T + \frac{1}{\alpha} I_n\right)\left(U_\gamma \Lambda_\gamma U_\gamma^T + \tilde{U}_\gamma \tilde{\Lambda}_\gamma \tilde{U}_\gamma^T\right) \\ &= U_\gamma U_\gamma^T + \frac{1}{\alpha} U_\gamma \Lambda_\gamma U_\gamma^T + \frac{1}{\alpha} \tilde{U}_\gamma \tilde{\Lambda}_\gamma \tilde{U}_\gamma^T \\ &= U_\gamma \left(I_n + \frac{1}{\alpha}\Lambda_\gamma\right) U_\gamma^T + \tilde{U}_\gamma \left(\frac{1}{\alpha}\tilde{\Lambda}_\gamma\right) \tilde{U}_\gamma^T, \end{aligned}$$

are

$$\begin{cases} \mu_i = 1 + \frac{\lambda_i}{\alpha} & \text{if} \quad \lambda_i \leq \gamma \quad (p \text{ eigenvalues}) \\ \mu_i = \frac{\lambda_i}{\alpha} & \text{if} \quad \lambda_i > \gamma \quad (n-p \text{ eigenvalues}) \end{cases} \tag{3.5}$$

and are included within the interval

$$\left[\min\left(\frac{\alpha + \lambda_{\min}(A)}{\alpha}, \frac{\gamma}{\alpha}\right), \max\left(\frac{\alpha+\gamma}{\alpha}, \frac{\lambda_{\max}(A)}{\alpha}\right)\right]. \tag{3.6}$$

Note that the condition number of $A_\gamma^{-1}A$ is explicitly controlled by the choice of the parameters $\alpha$ and $\gamma$ (with $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ fixed) and is given by

$$\kappa_2(A_\gamma^{-1}A) = \frac{\max\left(\frac{\alpha+\gamma}{\alpha}, \frac{\lambda_{\max}(A)}{\alpha}\right)}{\min\left(\frac{\alpha+\lambda_{\min}(A)}{\alpha}, \frac{\gamma}{\alpha}\right)}.$$

For instance, choosing $\gamma = \frac{\lambda_{\max}(A)}{100}$ and $\alpha = \frac{\lambda_{\max}(A)+\gamma}{2}$ yields

$$\begin{aligned} \frac{\gamma}{\alpha} &= \left(\frac{\lambda_{\max}(A)}{100}\right)\left(\frac{2}{\lambda_{\max}(A)+\gamma}\right) \\ &= \frac{2\lambda_{\max}(A)}{100\lambda_{\max}(A) + \lambda_{\max}(A)} \\ &= \frac{2}{101}, \end{aligned} \tag{3.7}$$

and

$$\begin{aligned} \frac{\lambda_{\max}(A)}{\alpha} &= \frac{2\lambda_{\max}(A)}{\lambda_{\max}(A)+\gamma} \\ &= \frac{2\lambda_{\max}(A)}{\lambda_{\max}(A) + \frac{\lambda_{\max}(A)}{100}} \\ &= \frac{200}{101}. \end{aligned} \tag{3.8}$$

Assuming that $\frac{\alpha+\lambda_{\min}(A)}{\alpha} = \mathcal{O}(1)$ and using (3.7) and (3.8) in (3.6), we obtain that the spectrum $\lambda(A_\gamma^{-1}A) = \{\mu_i\}_{i=1}^n$ of $A_\gamma^{-1}A$ satisfies,

$$\begin{aligned} \mu_i &\in \left[\min\left(\mathcal{O}(1), \frac{\gamma}{\alpha}\right), \max\left(1+\frac{\gamma}{\alpha}, \frac{\lambda_{\max}(A)}{\alpha}\right)\right] \\ &= \left[\min\left(\mathcal{O}(1), \frac{2}{101}\right), \max\left(1+\frac{2}{101}, \frac{200}{101}\right)\right] \\ &= \left[\frac{2}{101}, \frac{200}{101}\right], \end{aligned} \tag{3.9}$$

for $i = 1, \ldots, n$, so that

$$\kappa_2(A_\gamma^{-1}A) \leq 100.$$

Also observe that the matrix $A_\gamma^{-1}$ in (3.4) has similar ingredients to those used in deflation techniques but has not the form of the projector used in these techniques (see Giraud et al., 2006, for instance). Its effect is indeed not of a deflation type, in the sense that no eigenvalue of the matrix $A_\gamma^{-1}A$ is shifted to zero.

The choice of SLRU approach for the approximate inverse of $A$ is motivated by the simple form of the expression (3.4). The family of limited-memory preconditioners (LMP) could be a good alternative allowing other information as Ritz vectors or descent directions instead of spectral information to generalize the approach developed in this thesis. We refer to the PhD thesis of Tshimanga (2007) for more details on LMP preconditioners.

We illustrate the accuracy of the bounds (3.6) on a symmetric positive definite matrix of order $n = 300$ randomly generated by the `Matlab` function `sprandsym` (with a density of 0.05 and preset eigenvalues $\lambda_i = 10^{-8\times f_i}$, where $f_i$ is random uniformly distributed in $(0, 1)$). The smallest eigenvalue is $10^{-8}$, while the largest is $9.93\,10^{-1}$ implying that the condition number of $A$ is $9.9\,10^7$. As a first level of preconditioning, we consider the incomplete Cholesky decomposition of this matrix with a drop tolerance of $10^{-4}$ (see, e.g., Greenbaum, 1997, Section 11.1), followed by a Jacobi scaling to set the diagonal of the preconditioned matrix to 1. The spectrum of the resulting preconditioned matrix is well clustered, with 42 eigenvalues less than $\gamma = \frac{\lambda_{\max}(A)}{100} \approx 3.8\,10^{-2}$, and with extreme eigenvalues of $1.7\,10^{-7}$ and 3.8. The condition number is then $2.2\,10^7$. Figure 3.1 shows (on logarithmic scale) the eigenvalues of this preconditioned matrix. For simplicity, we shall denote as $A$ this preconditioned matrix in the following.

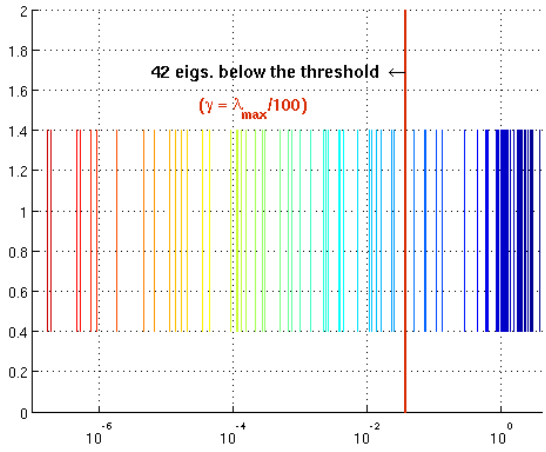



Figure 3.1 – Spectrum of the test matrix $A$ after incomplete Cholesky preconditioning and Jacobi scaling.

With the knowledge of the eigenvalues $\{\lambda_i\}_{i=1}^{42}$ (those less than $\gamma$) and the corresponding eigenvectors, we can set up $\Lambda_\gamma \in \mathbb{R}^{42\times 42}$ and $U_\gamma \in \mathbb{R}^{300\times 42}$ as well as $\alpha = 1.16$ (the average of the remaining eigenvalues). Finally, the bounds given by (3.6) ensure that the eigenvalues of $A_\gamma^{-1}A$ belong to $[3.2\,10^{-2}, 3.3]$ and $\kappa_2(A_\gamma^{-1}A) \leq 103.1$, which is illustrated by the eigenvalue distribution in Figure 3.2. The eigenvalue distribution of $A_\gamma^{-1}A$ shows a nice clustering around 1, which emphasizes the fact that $A_\gamma^{-1}$ is a good preconditioner for $A$.

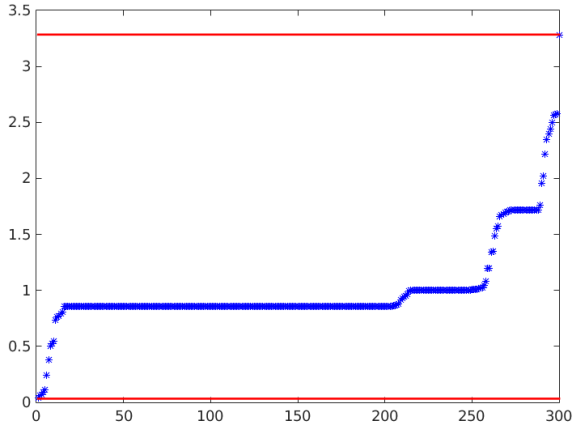



Figure 3.2 – Eigenvalue distribution of $A_\gamma^{-1}A$.

## 3.2 Spectral approximation of the Schur complement

We now introduce the approximation of the Schur complement of the matrices $\mathcal{A}_{KKT}$ and $\mathcal{A}_{SQD}$ as the matrices $S_\gamma = B^T A_\gamma^{-1} B$ and $S_\gamma = B^T A_\gamma^{-1} B + C$, respectively. Substituting (3.4) for $A_\gamma^{-1}$, the approximation $S_\gamma$ becomes

$$S_\gamma = B^T \left( U_\gamma \Lambda_\gamma^{-1} U_\gamma^T + \frac{1}{\alpha} I_n \right) B, \tag{3.10}$$

for the KKT case, and

$$S_\gamma = B^T A_\gamma^{-1} B + C = B^T \left( U_\gamma \Lambda_\gamma^{-1} U_\gamma^T + \frac{1}{\alpha} I_n \right) B + C, \tag{3.11}$$

for the SQD case. We refer the reader to the book of Zhang (2010) for a good survey on the Schur complement and its properties. In the next section, we develop an approximation for the inverse of $S_\gamma$ in (3.10) and we show that it is a good preconditioner for the Schur complement. Next, we generalize the theoretical result to the SQD case using (3.11) and point out the modifications implied on the bounds on the eigenvalues of the preconditioned Schur complement.

### 3.2.1 Spectral approximation of the inverse of the Schur complement for matrices of the KKT form

We write $S_\gamma$ in (3.10) using the notation

$$J_\gamma := B^T U_\gamma \Lambda_\gamma^{-\frac{1}{2}}, \tag{3.12}$$

as

$$\begin{aligned} S_\gamma &= (B^T U_\gamma \Lambda_\gamma^{-\frac{1}{2}})(\Lambda_\gamma^{-\frac{1}{2}} U_\gamma^T B) + \frac{1}{\alpha} B^T B \\ &= J_\gamma J_\gamma^T + \frac{1}{\alpha} B^T B. \end{aligned}$$

Then using the Sherman-Morrison-Woodbury formula (see Appendix A or Golub and Van Loan, 2013), we can derive the following expression for the inverse of $S_\gamma$,

$$S_\gamma^{-1} = \left( \frac{1}{\alpha} B^T B \right)^{-1} - \left( \frac{1}{\alpha} B^T B \right)^{-1} J_\gamma \left( I_n + J_\gamma^T \left( \frac{1}{\alpha} B^T B \right)^{-1} J_\gamma \right)^{-1} J_\gamma^T \left( \frac{1}{\alpha} B^T B \right)^{-1},$$

and substituting (3.12) for $J_\gamma$ yields,

$$\begin{aligned} S_\gamma^{-1} &= \alpha (B^T B)^{-1} \\ &\quad - \alpha^2 (B^T B)^{-1} B^T U_\gamma \left( \Lambda_\gamma + \alpha U_\gamma^T B (B^T B)^{-1} B^T U_\gamma \right)^{-1} U_\gamma^T B (B^T B)^{-1}, \end{aligned}$$

or, equivalently,

$$S_\gamma^{-1} = \alpha(B^TB)^{-1/2}\left(I_m - K_\gamma\left(\frac{1}{\alpha}\Lambda_\gamma + K_\gamma^TK_\gamma\right)^{-1}K_\gamma^T\right)(B^TB)^{-1/2}, \tag{3.13}$$

where $K_\gamma \in \mathbb{R}^{m\times p}$ is the operator defined by

$$K_\gamma := (B^TB)^{-1/2}B^TU_\gamma. \tag{3.14}$$

Observe that $K_\gamma$ involves the constraint matrix $B$ and the matrix $U_\gamma \in \mathbb{R}^{n\times p}$, which contains the orthonormal set of the $p$ eigenvectors associated to the eigenvalues in $A$ below a given threshold $\gamma$. The singular values of $K_\gamma$ correspond to the cosines of the principal angles between the two subspaces $\mathcal{I}m(B)$ and $\mathcal{I}m(U_\gamma)$, since $B(B^TB)^{-1/2}$ represents an orthonormal basis for $\mathcal{I}m(B)$ (see Appendix A or Golub and Van Loan, 2013, Section 6.4.3). The expression

$$K_\gamma\left(\frac{1}{\alpha}\Lambda_\gamma + K_\gamma^TK_\gamma\right)^{-1}K_\gamma^T$$

in (3.13) explicitly shows the interaction between $A$ and $B$, with the combined effects of both the smallest eigenvalues of $A$ and the cosines of the principal angles between $\mathcal{I}m(B)$ and $\mathcal{I}m(U_\gamma)$. This interaction between $A$ and $B$ will be studied in detail in Chapter 5. Note that the matrix $\frac{1}{\alpha}\Lambda_\gamma + K_\gamma^TK_\gamma \in \mathbb{R}^{p\times p}$ is a rank-$p$ update and is of a small dimension $p$.

Theorem 3.1 below gives the bounds that we have derived on the eigenvalues of the preconditioned Schur complement. This proof is based on the CS decomposition (see Appendix A or Paige and Saunders, 1981, Section 4). However, this result, submitted to the journal COAP[(2)] has been refined by an anonymous referee who gave us the proof, as shown in Theorem 3.2.

**Theorem 3.1** Let $A$ and $A_\gamma \in \mathbb{R}^{n\times n}$ be given by (3.3) and (3.4) respectively. Then the spectrum $\lambda(S_\gamma^{-1}S) = \{\nu_i\}_{i=1}^m$ of the matrix $S_\gamma^{-1}S \in \mathbb{R}^{m\times m}$ with $S = B^TA^{-1}B$ and $S_\gamma = B^TA_\gamma^{-1}B$ satisfies:

$$\nu_i \in \left[\frac{\alpha}{\alpha+\lambda_{\max}(A)+\gamma}, \frac{\alpha+\gamma}{\gamma}\right], \quad \text{for } i = 1,\ldots,m, \tag{3.15}$$

with $\alpha > 0$ and $\gamma \in [\lambda_{\min}(A), \lambda_{\max}(A)]$ as defined above.

(2) Computational Optimization and Applications.

**Proof.** See Appendix B for the proof. □

**Theorem 3.2** Let $A$ and $A_\gamma \in \mathbb{R}^{n\times n}$ be given by (3.3) and (3.4) respectively. Then the spectrum $\lambda(S_\gamma^{-1}S) = \{\nu_i\}_{i=1}^m$ of the matrix $S_\gamma^{-1}S \in \mathbb{R}^{m\times m}$ with $S = B^TA^{-1}B$ and $S_\gamma = B^TA_\gamma^{-1}B$ satisfies:

$$\nu_i \in \left[\min\left(\frac{\alpha}{\alpha+\gamma}, \frac{\alpha}{\lambda_{\max}(A)}\right), \max\left(\frac{\alpha}{\alpha+\lambda_{\min}(A)}, \frac{\alpha}{\gamma}\right)\right], \tag{3.16}$$

for $i = 1, \ldots, m$, with $\alpha > 0$ and $\gamma \in [\lambda_{\min}(A), \lambda_{\max}(A)]$.

**Proof.** First note that the matrices $S$ and $S_\gamma$ are symmetric and positive definite, hence nonsingular, by definition of $A$ and $A_\gamma$ respectively, and by the full column rank property of $B \in \mathbb{R}^{n\times m}$ (see, e.g., Golub and Van Loan, 2013, Section 4.2.1). The eigenvalue problem $S_\gamma^{-1}Sx = \lambda x$ is then equivalent to the generalized eigenvalue problem:

$$Sx = \lambda S_\gamma x, \tag{3.17}$$

that is, $\lambda(S_\gamma^{-1}S) = \lambda(S, S_\gamma) = \{\nu_i\}_{i=1}^m$.

Consider, for a non-zero vector $y \in \mathbb{R}^m$, the generalized Rayleigh quotient

$$\begin{aligned}\nu(y) &= \frac{y^TSy}{y^TS_\gamma y} \\ &= \frac{y^TB^TA^{-1}By}{y^TB^TA_\gamma^{-1}By},\end{aligned}$$

implying that the extreme eigenvalues of $S_\gamma^{-1}S$ are

$$\nu_m = \max_{y\in\mathbb{R}^m} \nu(y) \qquad \text{and} \qquad \nu_1 = \min_{y\in\mathbb{R}^m} \nu(y).$$

We first obtain the following inequality by the fact that the supremum of a function over $\mathbb{R}^n$ is greater than or equal to the supremum over the set of vectors of the form $x = By$, $y \in \mathbb{R}^m$,

$$\begin{aligned}\nu_m &\leq \max_{x\in\mathbb{R}^n} \frac{x^TA^{-1}x}{x^TA_\gamma^{-1}x} \\ &= \mu_{\max}(A_\gamma A^{-1}) \\ &= \frac{1}{\mu_{\min}(AA_\gamma^{-1})}.\end{aligned}$$

In the same way, we deduce that

$$\begin{aligned}\nu_1 &\geq \min_{x\in\mathbb{R}^n} \frac{x^T A^{-1} x}{x^T A_\gamma^{-1} x} \\ &= \mu_{\min}(A_\gamma A^{-1}) \\ &= \frac{1}{\mu_{\max}(A A_\gamma^{-1})}.\end{aligned}$$

By the interval (3.6) that contains the eigenvalues of $A_\gamma^{-1}A$, we conclude that

$$\begin{aligned}\nu_m &\leq \frac{1}{\min\left(\frac{\alpha+\lambda_{\min}(A)}{\alpha}, \frac{\gamma}{\alpha}\right)} \\ &= \max\left(\frac{\alpha}{\alpha+\lambda_{\min}(A)}, \frac{\alpha}{\gamma}\right)\end{aligned}$$

and

$$\begin{aligned}\nu_1 &\geq \frac{1}{\max\left(\frac{\alpha+\gamma}{\alpha}, \frac{\lambda_{\max}(A)}{\alpha}\right)} \\ &= \min\left(\frac{\alpha}{\alpha+\gamma}, \frac{\alpha}{\lambda_{\max}(A)}\right).\end{aligned}$$

□

The following theorem shows that the bounds on the eigenvalues of the preconditioned Schur complement given by Theorem 3.2 are sharper than the bounds given by Theorem 3.1.

**Theorem 3.3** The inequalities

$$\min\left(\frac{\alpha}{\alpha+\gamma}, \frac{\alpha}{\lambda_{\max}(A)}\right) \geq \frac{\alpha}{\alpha+\lambda_{\max}(A)+\gamma} \tag{3.18}$$

and

$$\max\left(\frac{\alpha}{\alpha+\lambda_{\min}(A)}, \frac{\alpha}{\gamma}\right) \leq \frac{\alpha+\gamma}{\gamma} \tag{3.19}$$

hold.

**Proof.** Since $\lambda_{\max}(A) > 0$, $\alpha > 0$ and $\gamma > 0$, we have

$$\frac{\alpha}{\alpha+\gamma} \geq \frac{\alpha}{\alpha+\lambda_{\max}(A)+\gamma} \qquad \text{and} \qquad \frac{\alpha}{\lambda_{\max}(A)} \geq \frac{\alpha}{\alpha+\lambda_{\max}(A)+\gamma},$$

implying the first inequality (3.18). In the same way, since

$$\frac{\alpha}{\alpha+\lambda_{\min}(A)} = 1 - \frac{\lambda_{\min}(A)}{\alpha+\lambda_{\min}(A)} < 1 \qquad \text{and} \qquad \frac{\alpha+\gamma}{\gamma} = \frac{\alpha}{\gamma} + 1 > 1,$$

as $\lambda_{\min}(A) > 0$, we obtain

$$\frac{\alpha}{\alpha+\lambda_{\min}(A)} \leq \frac{\alpha+\gamma}{\gamma}. \tag{3.20}$$

The inequality

$$\frac{\alpha}{\gamma} \leq \frac{\alpha+\gamma}{\gamma} \tag{3.21}$$

is obvious. By (3.20) and (3.21), we derive (3.19). □

Similarly to $A_\gamma^{-1}A$, the condition number of $S_\gamma^{-1}S$ is fully controlled by the choice of the parameters $\alpha$ and $\gamma$. For instance, choosing $\gamma = \frac{\lambda_{\max}(A)}{100}$ and $\alpha = \frac{\lambda_{\max}(A)+\gamma}{2}$, implies that

$$\begin{aligned} \alpha+\gamma &= \frac{\lambda_{\max}(A)+\gamma}{2} + \gamma \\ &= \frac{\lambda_{\max}(A)}{2} + \frac{3\gamma}{2} \\ &= \frac{103\lambda_{\max}(A)}{200} \end{aligned}$$

and by (3.7) and (3.8), we have

$$\frac{\alpha}{\gamma} = \frac{101}{2} \qquad \text{and} \qquad \frac{\alpha}{\lambda_{\max}(A)} = \frac{101}{200}. \tag{3.22}$$

Assuming again that $\dfrac{\alpha+\lambda_{\min}(A)}{\alpha} = \mathcal{O}(1)$, Theorem 3.2 implies that the spectrum $\lambda(S_\gamma^{-1}S) = \{\nu_i\}_{i=1}^m$ of the matrix $S_\gamma^{-1}S$ satisfies

$$\begin{aligned}
\nu_i &\in \left[\min\left(\frac{\frac{101}{200}\lambda_{\max}(A)}{\frac{103}{200}\lambda_{\max}(A)}, \frac{101}{200}\right), \max\left(\mathcal{O}(1), \frac{101}{2}\right)\right] \\
&= \left[\min\left(\frac{101}{103}, \frac{101}{200}\right), \max\left(\mathcal{O}(1), \frac{101}{2}\right)\right] \\
&= \left[\frac{101}{200}, \frac{101}{2}\right],
\end{aligned} \tag{3.23}$$

for $i = 1, \ldots, m$, so that

$$\kappa_2(S_\gamma^{-1} S) \leq 100.$$

We illustrate the tightness of the bounds in Theorem 3.2 using the previously introduced test example for the matrix $A$. The constraint matrix $B \in \mathbb{R}^{300\times 150}$ is built by means of the `Matlab` function `sprandn` (with a density of 0.05 and a condition number of $10^4$). The resulting Schur complement $S = B^T A^{-1} B$ is ill-conditioned, with a smallest eigenvalue of $2.4\,10^{-6}$ and a largest of $1.4\,10^5$. In the same way, the Schur complement approximation $S_\gamma = B^T A_\gamma^{-1} B$ exhibits the same ill-conditioning with extreme eigenvalues of $2.1\,10^{-6}$ and $1.4\,10^5$ and Figure 3.3 shows that the eigenvalue distribution of $S_\gamma$ is very close to the eigenvalue distribution of $S$. Finally, the bounds given by Theorem 3.2 ensure that the eigenvalues of $S_\gamma^{-1} S$ belong to $[0.3, 30.5]$ implying that $\kappa_2(S_\gamma^{-1} S) \leq 101.7$, which is illustrated by the eigenvalue distribution in Figure 3.4. With respect to these bounds, the extreme eigenvalues of $S_\gamma^{-1} S$ are $\nu_{\min} \approx 0.56$ and $\nu_{\max} \approx 16.28$. The eigenvalue distribution of $S_\gamma^{-1} S$ also shows a nice clustering around 1, which emphasizes the fact that $S_\gamma^{-1}$ is a good preconditioner for $S$.

### 3.2.2 Spectral approximation of the inverse of the Schur complement for matrices of the SQD form

In the previous section, we have introduced an approximation for the inverse of the Schur complement of KKT matrices and we have shown using one theoretical result that this approximation is a good preconditioner for the exact Schur complement. In this section, we extend the proposed approach for SQD matrices. Writing $S_\gamma$ in (3.11) using the notation (3.12), we get

$$\begin{aligned}
S_\gamma &= (B^T U_\gamma \Lambda_\gamma^{-\frac{1}{2}})(\Lambda_\gamma^{-\frac{1}{2}} U_\gamma^T B) + \frac{1}{\alpha} B^T B + C \\
&= J_\gamma J_\gamma^T + \frac{1}{\alpha} B^T B + C.
\end{aligned}$$

Using again the Sherman-Morrison-Woodbury formula, we can derive the following expression for the inverse of $S_\gamma$,

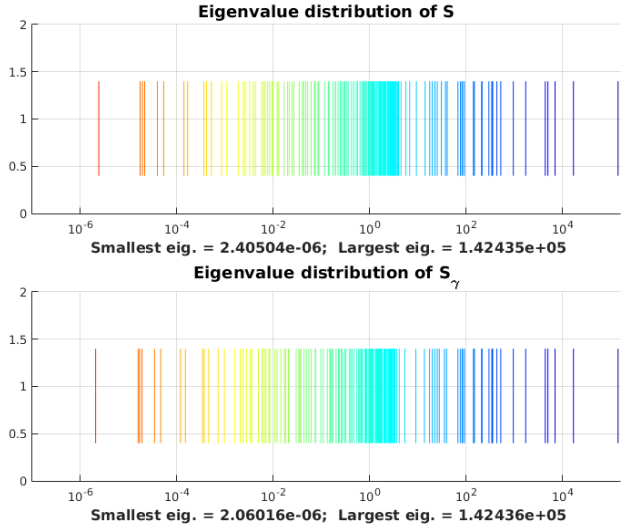



Figure 3.3 – Eigenvalue distribution of $S$ and $S_\gamma^{-1}$.

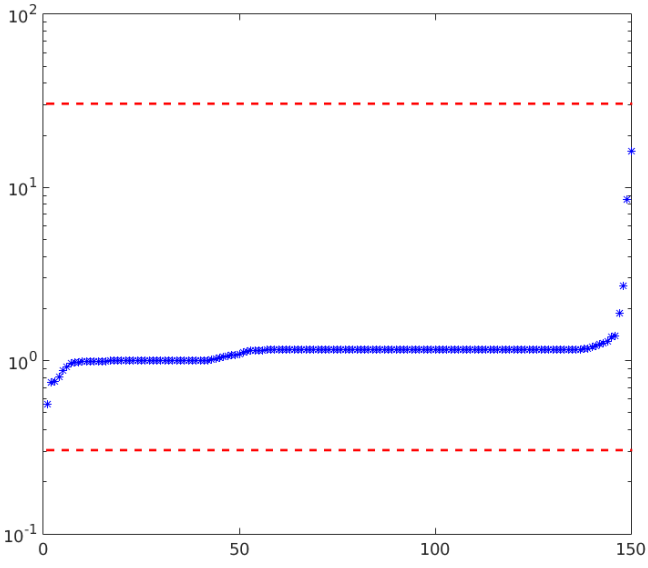


Figure 3.4 – Eigenvalue distribution of $S_\gamma^{-1}S$.

$$
\begin{aligned}
S_\gamma^{-1} &= \left(\frac{1}{\alpha}B^TB + C\right)^{-1} \\
&\quad - \left(\frac{1}{\alpha}B^TB + C\right)^{-1} J_\gamma \left(I_n + J_\gamma^T \left(\frac{1}{\alpha}B^TB + C\right)^{-1} J_\gamma\right)^{-1} J_\gamma^T \left(\frac{1}{\alpha}B^TB + C\right)^{-1},
\end{aligned}
$$

which yields, with the notation $B_\alpha = B^TB + \alpha C$,

$$
S_\gamma^{-1} = \alpha B_\alpha^{-1} - \alpha^2 B_\alpha^{-1} B^T U_\gamma (\Lambda_\gamma + \alpha U_\gamma^T B B_\alpha^{-1} B^T U_\gamma)^{-1} U_\gamma^T B B_\alpha^{-1},
$$

or, equivalently,

$$
S_\gamma^{-1} = \alpha (B^TB + \alpha C)^{-1/2} \left(I_m - K_\gamma \left(\frac{1}{\alpha}\Lambda_\gamma + K_\gamma^T K_\gamma\right)^{-1} K_\gamma^T\right) (B^TB + \alpha C)^{-1/2}, \tag{3.24}
$$

where $K_\gamma \in \mathbb{R}^{m\times p}$ is the operator defined by

$$
K_\gamma := (B^TB + \alpha C)^{-1/2} B^T U_\gamma. \tag{3.25}
$$

We point out the similarity of the matrix $S_\gamma^{-1}$ in (3.24) with the one associated to the KKT case (3.13). The only difference is that the term $B^TB$ is replaced by $B^TB + \alpha C$. In the same way, we have the following result, which establishes the lower and upper bounds on the eigenvalues of the Schur complement preconditioned by the matrix $S_\gamma^{-1}$ in (3.24).

**Theorem 3.4** Let $A$ and $A_\gamma \in \mathbb{R}^{n\times n}$ be given by (3.3) and (3.4), respectively. Then the spectrum $\lambda(S_\gamma^{-1}S) = \{\nu_i\}_{i=1}^m$ of the matrix $S_\gamma^{-1}S \in \mathbb{R}^{m\times m}$ with $S = B^TA^{-1}B + C$ and $S_\gamma = B^TA_\gamma^{-1}B + C$ satisfies:

$$
\nu_i \in \left[\frac{\alpha}{2\alpha + \lambda_{\max}(A) + \gamma}, \frac{\alpha + 2\gamma}{\gamma}\right], \quad \text{for } i = 1, \ldots, m, \tag{3.26}
$$

with $\alpha > 0$ and $\gamma \in [\lambda_{\min}(A), \lambda_{\max}(A)]$.

**Proof.** The proof follows steps similar to the ones of Theorem 3.1 and uses the positive definiteness of $C$, see Appendix B. □

We have also generalized the proof of Theorem 3.2 given by the anonymous referee.

**Theorem 3.5** Let $A$ and $A_\gamma \in \mathbb{R}^{n\times n}$ be given by (3.3) and (3.4), respectively. Then the spectrum $\lambda(S_\gamma^{-1}S) = \{\nu_i\}_{i=1}^m$ of the matrix $S_\gamma^{-1}S \in \mathbb{R}^{m\times m}$ with $S = B^TA^{-1}B + C$ and $S_\gamma = B^TA_\gamma^{-1}B + C$ satisfies,

$$\nu_i \in \left[\min\left(\frac{\alpha}{2\alpha+\gamma}, \frac{\alpha}{\alpha+\lambda_{\max}(A)}\right), \max\left(\frac{2\alpha+\lambda_{\min}(A)}{\alpha+\lambda_{\min}(A)}, \frac{\alpha+\gamma}{\gamma}\right)\right], \quad (3.27)$$

for $i = 1, \ldots, m$, with $\alpha > 0$ and $\gamma \in [\lambda_{\min}(A), \lambda_{\max}(A)]$.

**Proof.** The matrices $S$ and $S_\gamma$ are symmetric and positive definite, hence nonsingular, by definition of $A$, $A_\gamma$ and $C$, and by the full column rank property of $B \in \mathbb{R}^{n\times m}$ (see, e.g., Golub and Van Loan, 2013, Section 4.2.1). We now consider the eigenvalue problem $S_\gamma^{-1}Sx = \lambda x$, which is equivalent to the generalized eigenvalue problem:

$$Sx = \lambda S_\gamma x, \quad (3.28)$$

that is, $\lambda(S_\gamma^{-1}S) = \lambda(S, S_\gamma) = \{\nu_i\}_{i=1}^m$.

The generalized Rayleigh quotient, for a non-zero vector $y \in \mathbb{R}^m$, is defined as

$$\nu(y) = \frac{y^TSy}{y^TS_\gamma y}. \quad (3.29)$$

Using the definitions of $S$ and $S_\gamma$, one can thus write

$$\begin{aligned}\nu(y) &= \frac{y^T(B^TA^{-1}B + C)y}{y^T(B^TA_\gamma^{-1}B + C)y} \\ &= \frac{y^TB^TA^{-1}By}{y^T(B^TA_\gamma^{-1}B + C)y} + \frac{y^TCy}{y^T(B^TA_\gamma^{-1}B + C)y}, \end{aligned} \quad (3.30)$$

and in the same way,

$$\begin{aligned}\frac{1}{\nu(y)} &= \frac{y^T(B^TA_\gamma^{-1}B + C)y}{y^T(B^TA^{-1}B + C)y} \\ &= \frac{y^TB^TA_\gamma^{-1}By}{y^T(B^TA^{-1}B + C)y} + \frac{y^TCy}{y^T(B^TA^{-1}B + C)y}. \end{aligned} \quad (3.31)$$

Since the matrices $C$, $B^TA^{-1}B$ and $B^TA_\gamma^{-1}B$ are symmetric positive definite, we have

$$\nu(y) \leq \frac{y^TB^TA^{-1}By}{y^TB^TA_\gamma^{-1}By} + 1 \quad (3.32)$$

and

$$\frac{1}{\nu(y)} \leq \frac{y^T B^T A_\gamma^{-1} B y}{y^T B^T A^{-1} B y} + 1. \tag{3.33}$$

The fact that the supremum of a function over $\mathbb{R}^n$ is greater than or equal to the supremum over the set of vectors of the form $x = By$ with $y \in \mathbb{R}^m$, implies, using (3.32),

$$\begin{aligned} \nu_m = \max_{y \in \mathbb{R}^m} \nu(y) &\leq \max_{y \in \mathbb{R}^m} \frac{y^T B^T A^{-1} B y}{y^T B^T A_\gamma^{-1} B y} + 1 \\ &\leq \max_{x \in \mathbb{R}^n} \frac{x^T A^{-1} x}{x^T A_\gamma^{-1} x} + 1 \\ &= \mu_{\max}(A_\gamma A^{-1}) + 1. \\ &= \frac{1}{\mu_{\min}(A A_\gamma^{-1})} + 1. \end{aligned} \tag{3.34}$$

Similarly, we have, by (3.33),

$$\begin{aligned} \nu_1 = \min_{y \in \mathbb{R}^m} \nu(y) &\geq \min_{y \in \mathbb{R}^m} \frac{1}{\frac{y^T B^T A_\gamma^{-1} B y}{y^T B^T A^{-1} B y} + 1} \\ &\geq \min_{x \in \mathbb{R}^n} \frac{1}{\frac{x^T A_\gamma^{-1} x}{x^T A^{-1} x} + 1} \\ &= \frac{1}{\max\limits_{x \in \mathbb{R}^n} \frac{x^T A_\gamma^{-1} x}{x^T A^{-1} x} + 1} \\ &= \frac{1}{\mu_{\max}(A A_\gamma^{-1}) + 1}. \end{aligned} \tag{3.35}$$

Using (3.6), we can deduce that

$$\begin{aligned} \nu_m &\leq \frac{1}{\min\left(\frac{\alpha + \lambda_{\min}(A)}{\alpha}, \frac{\gamma}{\alpha}\right)} + 1 \\ &= \max\left(\frac{\alpha}{\alpha + \lambda_{\min}(A)}, \frac{\alpha}{\gamma}\right) + 1 \\ &= \max\left(\frac{2\alpha + \lambda_{\min}(A)}{\alpha + \lambda_{\min}(A)}, \frac{\alpha + \gamma}{\gamma}\right) \end{aligned}$$

and

$$\begin{aligned}\nu_1 &\geq \frac{1}{\max\left(\frac{\alpha+\gamma}{\alpha}, \frac{\lambda_{\max}(A)}{\alpha}\right)+1} \\ &= \frac{1}{\max\left(\frac{2\alpha+\gamma}{\alpha}, \frac{\alpha+\lambda_{\max}(A)}{\alpha}\right)} \\ &= \min\left(\frac{\alpha}{2\alpha+\gamma}, \frac{\alpha}{\alpha+\lambda_{\max}(A)}\right).\end{aligned}$$

□

We can again illustrate the tightness of the bounds in Theorem 3.5 on the previously introduced test example for the matrices $A$ and $B$ (see Sections 3.1 and 3.2.1). The matrix $C \in \mathbb{R}^{150\times 150}$ is built by means of the `Matlab` function `sprandn` (with a density of 0.05 and a condition number of $10^4$). The bounds given by Theorem 3.5 ensure that the eigenvalues of $S_\gamma^{-1}S$ belong to $[0.2, 31.5]$, which is illustrated by the eigenvalue distribution in Figure 3.5. With respect to these bounds, the extreme eigenvalues of $S_\gamma^{-1}S$ are $\nu_{\min} \approx 0.62$ and $\nu_{\max} \approx 9.76$. The eigenvalue distribution of $S_\gamma^{-1}S$ also shows a nice clustering around 1, which emphasizes the fact that $S_\gamma^{-1}$ is a good preconditioner for $S$.

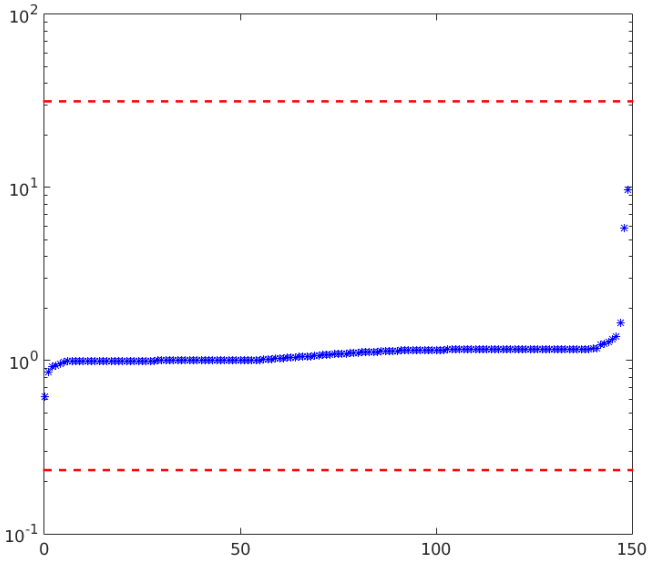



Figure 3.5 – Eigenvalue distribution of $S_\gamma^{-1}S$.

Chapter 4

# Spectral preconditioners for saddle-point matrices

In this chapter, we investigate two alternatives to get efficient approximations of the "ideal" block diagonal preconditioner

$$\mathcal{P} = \begin{bmatrix} A & 0 \\ 0 & S \end{bmatrix}, \tag{4.1}$$

with the Schur complement $S = B^T A^{-1} B$ proposed by Murphy et al. (2000) for the KKT matrix

$$\mathcal{A}_{KKT} = \begin{bmatrix} A & B \\ B^T & 0 \end{bmatrix}, \tag{4.2}$$

and with $S = B^T A^{-1} B + C$ proposed by Gould and Simoncini (2009) for the SQD matrix

$$\mathcal{A}_{SQD} = \begin{bmatrix} A & B \\ B^T & -C \end{bmatrix}. \tag{4.3}$$

As we have seen in Chapter 2, the performance of the MINRES method depends on the distribution of the eigenvalues of the saddle-point matrix, which are bounded within the intervals given in Theorems 2.5 and 2.6 for the KKT matrix and the SQD matrix, respectively.

The preconditioner in (4.1) may be computationally expensive and in practice, approximations of $A$ and of the Schur complement are necessary.

In Pestana and Wathen (2014), the authors study the bounds on the eigenvalues of saddle-point systems preconditioned by block diagonal preconditioners when saddle-point systems require discretization as for instance, electromagnetic problems or incompressible fluid dynamics problems. With respect to block diagonal preconditioners of the form (4.1) combining the knowledge of some spectral information, we refer the reader to Olshanskii and Simoncini

(2010). In this work, the authors perform an analysis of the eigenvalue distribution of the preconditioned Schur complement matrix, showing how the presence of a few outliers in this preconditioned Schur complement matrix is accurately inherited by the global preconditioned matrix $\mathcal{P}^{-1}\mathcal{A}_{KKT}$. They then propose a strategy to accelerate the convergence of MINRES with a deflation technique according to which they incorporate an approximation of those eigenvectors into the preconditioned matrix $\mathcal{P}^{-1}\mathcal{A}_{KKT}$ associated with the (inherited) outlying eigenvalues closest to zero.

Following the developments in Chapter 3, we incorporate the approximations $A_\gamma^{-1}$ and $S_\gamma^{-1}$ for the inverse of $A$ and $S$ for the KKT systems or the SQD systems, in Sections 4.1 and 4.2 respectively, to approximate the inverse of $\mathcal{P}$. These sections introduce each two alternatives, of the form

$$\mathcal{P}_1 := \begin{bmatrix} A_\gamma & 0 \\ 0 & S_\gamma \end{bmatrix} \quad \text{and} \quad \mathcal{P}_2 := \begin{bmatrix} A & 0 \\ 0 & S_\gamma \end{bmatrix},$$

for the inverse of $\mathcal{P}$ and we give the theoretical bounds on the eigenvalues of the preconditioned KKT matrices or SQD matrices, respectively. Next, we compare the effectiveness of these alternative block diagonal preconditioners on KKT systems and SQD systems in Section 4.3 and 4.4, respectively. Finally, we focus on the KKT systems and we combine the preconditioners with a first level of preconditioning.

## 4.1 Spectral preconditioners for the KKT systems

To simplify the writing, we make the short notation $\mathcal{P}_1$ to denote our first alternative for $\mathcal{P}$ in (4.1), which we approximate with

$$\mathcal{P}_1 := \begin{bmatrix} A_\gamma & 0 \\ 0 & S_\gamma \end{bmatrix}, \tag{4.4}$$

where $S_\gamma = B^T A_\gamma^{-1} B$ with $A_\gamma^{-1} = U_\gamma \Lambda_\gamma^{-1} U_\gamma^T + \frac{1}{\alpha} I_n$ as given in (3.4). The following theorem gives the bounds on the eigenvalues of a $\mathcal{A}_{KKT}$ preconditioned by $\mathcal{P}_1$.

**Theorem 4.1** Let $\mathcal{P}_1$ be given by (4.4) with $A_\gamma^{-1} = U_\gamma \Lambda_\gamma^{-1} U_\gamma^T + \frac{1}{\alpha} I_n$ and $S_\gamma = B^T A_\gamma^{-1} B$. Then the eigenvalues of the preconditioned matrix $\mathcal{P}_1^{-1}\mathcal{A}_{KKT}$ are bounded within the intervals

$$\left[\frac{\mu_{\min} - \sqrt{\mu_{\min}^2 + 4}}{2}, \frac{\mu_{\max} - \sqrt{\mu_{\max}^2 + 4}}{2}\right] \bigcup \left[\mu_{\min}, \frac{\mu_{\max} + \sqrt{\mu_{\max}^2 + 4}}{2}\right]$$

where

$$\mu_{\min} = \min\left(\frac{\alpha + \lambda_{\min}(A)}{\alpha}, \frac{\gamma}{\alpha}\right) \quad \text{and} \quad \mu_{\max} = \max\left(\frac{\alpha + \gamma}{\alpha}, \frac{\lambda_{\max}(A)}{\alpha}\right) \tag{4.5}$$

denote the lower and upper bounds on the eigenvalues of $A_\gamma^{-1}A$ given by (3.6).

**Proof.** Note that $\mathcal{P}_1^{-1}\mathcal{A}_{KKT}$ is similar to $\mathcal{P}_1^{-1/2}\mathcal{A}_{KKT}\mathcal{P}_1^{-1/2}$, with

$$\begin{aligned} \mathcal{P}_1^{-1/2}\mathcal{A}_{KKT}\mathcal{P}_1^{-1/2} &= \begin{bmatrix} A_\gamma^{-1/2} & 0 \\ 0 & S_\gamma^{-1/2} \end{bmatrix} \begin{bmatrix} A & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} A_\gamma^{-1/2} & 0 \\ 0 & S_\gamma^{-1/2} \end{bmatrix} \\ &= \begin{bmatrix} A_\gamma^{-1/2} A A_\gamma^{-1/2} & Q_1 \\ Q_1^T & 0 \end{bmatrix}, \end{aligned} \tag{4.6}$$

where $Q_1 = A_\gamma^{-1/2} B S_\gamma^{-1/2}$ satisfies

$$\begin{aligned} Q_1^T Q_1 &= S_\gamma^{-1/2} B^T A_\gamma^{-1/2} A_\gamma^{-1/2} B S_\gamma^{-1/2} \\ &= S_\gamma^{-1/2} S_\gamma S_\gamma^{-1/2} \\ &= I_m. \end{aligned} \tag{4.7}$$

We then recall that the eigenvalues of $A_\gamma^{-1}A$ (which is similar to $A_\gamma^{-1/2} A A_\gamma^{-1/2}$ ) are bounded within the interval $[\mu_{\min}, \mu_{\max}]$, with $\mu_{\min}$ and $\mu_{\max}$ defined by (4.5). Observing that the singular values $\sigma_i$ of $Q_1$ in (4.6) are equal to 1 by (4.7), we get the desired result from the bounds in Theorem 2.5 applied on (4.6) with $\sigma_1 = \sigma_m = 1$, $\lambda_1 = \mu_{\min}$ and $\lambda_n = \mu_{\max}$. □

For instance, choosing as previously in Section 3.1 $\gamma = \frac{\lambda_{\max}(A)}{100}$ and $\alpha = \frac{\lambda_{\max}(A)+\gamma}{2}$ yields

$$\mu_{\min} = \frac{2}{101} \quad \text{and} \quad \mu_{\max} = \frac{200}{101}$$

as computed in (3.9) and by Theorem 4.1, the eigenvalues of $\mathcal{P}_1^{-1}\mathcal{A}_{KKT}$ are included within the interval

$$\left[\frac{\frac{2}{101}-\sqrt{\left(\frac{2}{101}\right)^2+4}}{2},\frac{\frac{200}{101}-\sqrt{\left(\frac{200}{101}\right)^2+4}}{2}\right]\bigcup\left[\frac{2}{101},\frac{\frac{200}{101}+\sqrt{\left(\frac{200}{101}\right)^2+4}}{2}\right],$$

or, equivalently, $[-0.99,-0.42]\cup[0.02,2.40]$, and the condition number of preconditioned matrix $\kappa_2(\mathcal{P}_1^{-1}\mathcal{A}_{KKT}) \leq 121.21$.

Observe that, by Theorem 4.1, the left interval, associated to the negative eigenvalues in $\mathcal{P}_1^{-1}\mathcal{A}_{KKT}$, is basically well bounded and isolated away from zero, as opposed to the right interval (the one associated to the positive eigenvalues in $\mathcal{P}_1^{-1}\mathcal{A}_{KKT}$), which is well bounded towards infinity but not isolated away from zero. Indeed, we can assume from (4.5), that $\mu_{\max} = \mathcal{O}(1)$ (for reasonable choices of $\alpha$), whereas $\mu_{\min} = \min(\mathcal{O}(1), \mathcal{O}(\gamma))$ can actually tend to zero with small values for the parameter $\gamma$, and this mostly influences only the lower bound in the right interval.

The other alternative to approximate the "ideal" block diagonal preconditioner $\mathcal{P}$ is to replace only the Schur complement by its approximation $S_\gamma$. Indeed, knowing that the spectral information extracted from the $(1,1)$ block $A$ is readily available, it is also reasonable to consider that solutions with $A$ can be obtained in a cheap way by means of deflated Krylov techniques as in Giraud et al. (2006). We thus consider the second preconditioner

$$\mathcal{P}_2 := \begin{bmatrix} A & 0 \\ 0 & S_\gamma \end{bmatrix} \tag{4.8}$$

and we obtain the following result about the clustering of the eigenvalues in the preconditioned matrix $\mathcal{P}_2^{-1}\mathcal{A}_{KKT}$, which uses the result below on the eigenvalues of matrix where the $(1,1)$ block is the identity.

**Theorem 4.2** Let $0 < \sigma_1 \leq \sigma_2 \leq \ldots \leq \sigma_m$ the singular values of $B$. Then the eigenvalues of

$$\begin{bmatrix} I_n & B \\ B^T & 0 \end{bmatrix}$$

are $0$ with multiplicity $r$, $1$ with multiplicity $n-m+r$ and $\frac{1\pm\sqrt{1+4\sigma_i^2}}{2}$ for $i = 1,\ldots,m-r$.

**Proof.** (Fischer, Ramage, Silvester and Wathen, 1998[(1)], Lemma 2.1 for $\eta = 1$). Note that, when $B$ has a full column rank, the KKT matrix is nonsingular

[(1)] This reference has been given by an anonymous referee.

implying that $r = 0$. □

We can now state the following theorem, which gives the bounds on the eigenvalues of $\mathcal{P}_2^{-1}\mathcal{A}_{KKT}$.

**Theorem 4.3** Let $\mathcal{P}_2$ be given by (4.8) with $A_\gamma^{-1} = U_\gamma \Lambda_\gamma^{-1} U_\gamma^T + \frac{1}{\alpha} I_n$ and $S_\gamma = B^T A_\gamma^{-1} B$. Then the eigenvalues of the preconditioned matrix $\mathcal{P}_2^{-1}\mathcal{A}_{KKT}$ are bounded within the intervals

$$\left[\frac{1-\sqrt{1+4\nu_{\max}}}{2}, \frac{1-\sqrt{1+4\nu_{\min}}}{2}\right] \cup \{1\} \cup \left[\frac{1+\sqrt{1+4\nu_{\min}}}{2}, \frac{1+\sqrt{1+4\nu_{\max}}}{2}\right]$$

where

$$\nu_{\min} = \min\left(\frac{\alpha}{\alpha+\gamma}, \frac{\alpha}{\lambda_{\max}(A)}\right) \quad \text{and} \quad \nu_{\max} = \max\left(\frac{\alpha}{\alpha+\lambda_{\min}(A)}, \frac{\alpha}{\gamma}\right) \tag{4.9}$$

denote the lower and upper bounds on the eigenvalues of $S_\gamma^{-1} S$ given by (3.16).

**Proof.** Note that $\mathcal{P}_2^{-1}\mathcal{A}_{KKT}$ is similar to $\mathcal{P}_2^{-1/2}\mathcal{A}_{KKT}\mathcal{P}_2^{-1/2}$, with

$$\begin{aligned} \mathcal{P}_2^{-1/2}\mathcal{A}_{KKT}\mathcal{P}_2^{-1/2} &= \begin{bmatrix} A^{-1/2} & 0 \\ 0 & S_\gamma^{-1/2} \end{bmatrix} \begin{bmatrix} A & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} A^{-1/2} & 0 \\ 0 & S_\gamma^{-1/2} \end{bmatrix} \\ &= \begin{bmatrix} I_n & Q_2 \\ Q_2^T & 0 \end{bmatrix}, \end{aligned} \tag{4.10}$$

where $Q_2 = A^{-1/2} B S_\gamma^{-1/2}$ satisfies

$$\begin{aligned} Q_2^T Q_2 &= S_\gamma^{-1/2} B^T A^{-1/2} A^{-1/2} B S_\gamma^{-1/2} \\ &= S_\gamma^{-1/2} S S_\gamma^{-1/2}. \end{aligned}$$

As we have seen, the eigenvalues $\nu_i$ of $S_\gamma^{-1} S$ (which is similar to $S_\gamma^{-1/2} S S_\gamma^{-1/2}$) are bounded within the interval $[\nu_{\min}, \nu_{\max}]$, with $\nu_{\min}$ and $\nu_{\max}$ defined by (4.9). We deduce from Theorem 4.2 applied on (4.10), that the eigenvalues of (4.10) are 1 with multiplicity $n - m$ and $\frac{1 \pm \sqrt{1+4\sigma_i^2}}{2}$ for $i = 1, \ldots, m$ implying the bounds of the desired result. □

For instance, choosing as previously in Section 3.1 $\gamma = \frac{\lambda_{\max}(A)}{100}$ and $\alpha = \frac{\lambda_{\max}(A)+\gamma}{2}$ yields

$$\nu_{\min} = \frac{101}{200} \quad \text{and} \quad \nu_{\max} = \frac{101}{2}$$

as computed in (3.23), so that Theorem 4.3 implies that the eigenvalues of $\mathcal{P}_2^{-1}\mathcal{A}_{KKT}$ are included within the interval

$$\left[\frac{1-\sqrt{1+4\left(\frac{101}{2}\right)}}{2}, \frac{1-\sqrt{1+4\left(\frac{101}{200}\right)}}{2}\right] \cup \{1\} \cup \left[\frac{1+\sqrt{1+4\left(\frac{101}{200}\right)}}{2}, \frac{1+\sqrt{1+4\left(\frac{101}{2}\right)}}{2}\right]$$

or, equivalently, $[-6.62, -0.37] \cup \{1\} \cup [1.37, 7.62]$, and the condition number $\kappa_2(\mathcal{P}_2^{-1}\mathcal{A}_{KKT}) \leq 20.59$.

Note that Theorem 4.3 allows isolating the eigenvalue 1 and tightening the bounds on the positive part of the spectrum. This can be of interest when deriving upper bounds for the rate of convergence in Krylov methods such as MINRES for instance. Indeed, it is possible to partly refine the convergence rate by incorporating the specific root 1 into the polynomials used to establish this rate, and obtain a rate that depends directly on the bounds of the two extreme intervals (without taking the value 1 into account).

## 4.2 Spectral preconditioners for the SQD systems

Similarly to the previous section, we now use Theorem 2.6 to generalize the previous results to the SQD matrix. To simplify the writing, we make again the short notation[(2)] $\mathcal{P}_1$ to denote our first alternative for $\mathcal{P}$, which we approximate with

$$\mathcal{P}_1 := \begin{bmatrix} A_\gamma & 0 \\ 0 & S_\gamma \end{bmatrix} \tag{4.11}$$

where the Schur complement $S_\gamma = B^T A_\gamma^{-1} B + C$ with $A_\gamma^{-1} = U_\gamma \Lambda_\gamma^{-1} U_\gamma^T + \frac{1}{\alpha} I_n$. The next result gives the bounds on the eigenvalues of a SQD matrix preconditioned by $\mathcal{P}_1$.

[(2)] We draw attention on the fact that we use the same notation for both alternatives of the preconditioners in the case of the KKT systems and the SQD systems.

**Theorem 4.4** Let $\mathcal{P}_1$ be given by (4.11) with $A_\gamma^{-1} = U_\gamma \Lambda_\gamma^{-1} U_\gamma^T + \frac{1}{\alpha} I_n$ and $S_\gamma = B^T A_\gamma^{-1} B + C$. Then the eigenvalues of the preconditioned matrix $\mathcal{P}_1^{-1}\mathcal{A}_{SQD}$ are bounded within the intervals

$$\left[\frac{\mu_{\min} - \omega_{\max} - \sqrt{(\mu_{\min} + \omega_{\max})^2 + 4\bar{\omega}_{\max}}}{2}, \frac{\mu_{\max} - \sqrt{\mu_{\max}^2 + 4\bar{\omega}_{\min}}}{2}\right]$$

$$\bigcup$$

$$\left[\mu_{\min}, \frac{\mu_{\max} + \sqrt{\mu_{\max}^2 + 4\bar{\omega}_{\max}}}{2}\right]$$

where

$$\mu_{\min} = \min\left(\frac{\alpha + \lambda_{\min}(A)}{\alpha}, \frac{\gamma}{\alpha}\right) \quad \text{and} \quad \mu_{\max} = \max\left(\frac{\alpha + \gamma}{\alpha}, \frac{\lambda_{\max}(A)}{\alpha}\right) \tag{4.12}$$

denote the lower and upper bounds on the eigenvalues of $A_\gamma^{-1}A$ given by (3.6), with $\bar{\omega}_{\min}$ and $\bar{\omega}_{\max}$, the smallest and the largest eigenvalues of $S_\gamma^{-1}(B^T A_\gamma^{-1} B)$, respectively and with $\omega_{\max}$ the largest eigenvalues of $S_\gamma^{-1}C$.

**Proof.** Note that $\mathcal{P}_1^{-1}\mathcal{A}_{SQD}$ is similar to $\mathcal{P}_1^{-1/2}\mathcal{A}_{SQD}\mathcal{P}_1^{-1/2}$, with

$$\begin{aligned}
\mathcal{P}_1^{-1/2}\mathcal{A}_{SQD}\mathcal{P}_1^{-1/2} &= \begin{bmatrix} A_\gamma^{-1/2} & 0 \\ 0 & S_\gamma^{-1/2} \end{bmatrix} \begin{bmatrix} A & B \\ B^T & -C \end{bmatrix} \begin{bmatrix} A_\gamma^{-1/2} & 0 \\ 0 & S_\gamma^{-1/2} \end{bmatrix} \\
&= \begin{bmatrix} A_\gamma^{-1/2} A A_\gamma^{-1/2} & Q_1 \\ Q_1^T & -S_\gamma^{-1/2} C S_\gamma^{-1/2} \end{bmatrix},
\end{aligned} \tag{4.13}$$

where $Q_1 = A_\gamma^{-1/2} B S_\gamma^{-1/2}$ satisfies

$$\begin{aligned}
Q_1^T Q_1 &= S_\gamma^{-1/2} B^T A_\gamma^{-1/2} A_\gamma^{-1/2} B S_\gamma^{-1/2} \\
&= S_\gamma^{-1/2} B^T A_\gamma^{-1} B S_\gamma^{-1/2}
\end{aligned}$$

The matrix $S_\gamma^{-1/2} B^T A_\gamma^{-1} B S_\gamma^{-1/2}$ is similar to $S_\gamma^{-1}(B^T A_\gamma^{-1} B)$ and we get the desired result from the bounds in Theorem 2.6 applied on (4.13) with $\lambda_1 = \mu_{\min}$, $\lambda_n = \mu_{\max}$, $\sigma_1^2 = \bar{\omega}_{\min}$, $\sigma_m^2 = \bar{\omega}_{\max}$ and $\lambda_m^C = \omega_{\max}$. □

We consider now the alternative preconditioner for SQD systems,

$$\mathcal{P}_2 := \begin{bmatrix} A & 0 \\ 0 & S_\gamma \end{bmatrix} \tag{4.14}$$

for which we have the following result about the clustering of the eigenvalues in the preconditioned matrix $\mathcal{P}_2^{-1}\mathcal{A}_{SQD}$.

**Theorem 4.5** Let $\mathcal{P}_2$ be given by (4.14) with $A_\gamma^{-1} = U_\gamma \Lambda_\gamma^{-1} U_\gamma^T + \frac{1}{\alpha} I_n$ and $S_\gamma = B^T A_\gamma^{-1} B + C$. Then the eigenvalues of the preconditioned matrix $\mathcal{P}_2^{-1}\mathcal{A}_{SQD}$ are bounded within the intervals

$$\left[ \frac{1 - \omega_{\max} - \sqrt{(1+\omega_{\max})^2 + 4\hat{\omega}_{\max}}}{2}, \frac{1 - \sqrt{1 + 4\hat{\omega}_{\min}}}{2} \right]$$

$$\bigcup$$

$$\left[ 1, \frac{1 + \sqrt{1 + 4\hat{\omega}_{\max}}}{2} \right]$$

where

$$\nu_{\min} = \min\left( \frac{\alpha}{2\alpha+\gamma}, \frac{\alpha}{\alpha + \lambda_{\max}(A)} \right) \quad \text{and} \quad \nu_{\max} = \max\left( \frac{2\alpha + \lambda_{\min}(A)}{\alpha + \lambda_{\min}(A)}, \frac{\alpha+\gamma}{\gamma} \right) \tag{4.15}$$

denote the lower and upper bounds on the eigenvalues of $S_\gamma^{-1} S$ given by (3.27), with $\hat{\omega}_{\min}$ and $\hat{\omega}_{\max}$, the smallest and the largest eigenvalues of $S_\gamma^{-1}(B^T A^{-1} B)$, respectively and with $\omega_{\max}$ the largest eigenvalues of $S_\gamma^{-1} C$.

**Proof.** Note that $\mathcal{P}_2^{-1}\mathcal{A}_{SQD}$ is similar to $\mathcal{P}_2^{-1/2}\mathcal{A}_{SQD}\mathcal{P}_2^{-1/2}$, with

$$\begin{aligned} \mathcal{P}_2^{-1/2}\mathcal{A}_{SQD}\mathcal{P}_2^{-1/2} &= \begin{bmatrix} A^{-1/2} & 0 \\ 0 & S_\gamma^{-1/2} \end{bmatrix} \begin{bmatrix} A & B \\ B^T & -C \end{bmatrix} \begin{bmatrix} A^{-1/2} & 0 \\ 0 & S_\gamma^{-1/2} \end{bmatrix} \\ &= \begin{bmatrix} I_n & Q_2 \\ Q_2^T & -S_\gamma^{-1/2} C S_\gamma^{-1/2} \end{bmatrix}, \end{aligned} \tag{4.16}$$

where $Q_2 = A^{-1/2} B S_\gamma^{-1/2}$ satisfies

$$\begin{aligned} Q_2^T Q_2 &= S_\gamma^{-1/2} B^T A^{-1/2} A^{-1/2} B S_\gamma^{-1/2} \\ &= S_\gamma^{-1/2} B^T A^{-1} B S_\gamma^{-1/2} \end{aligned}$$

The matrix $S_\gamma^{-1/2} B^T A^{-1} B S_\gamma^{-1/2}$ is similar to $S_\gamma^{-1}(B^T A^{-1} B)$ and we deduce from the bounds in Theorem 2.6 applied on (4.16) with $\sigma_1^2 = \hat{\omega}_{\min}$, $\sigma_m^2 = \hat{\omega}_{\max}$ and $\lambda_m^C = \omega_{\max}$ the desired result. □

## 4.3 Comparison of the spectral preconditioners for the KKT systems

In this section, we illustrate and compare the effectiveness of the two preconditioning alternatives presented in Section 4.1 on the previously introduced test example (see Section 3.1) for the KKT systems. We provide, in Table 4.1, the true negative and positive intervals in which the eigenvalues of $\mathcal{P}_1^{-1}\mathcal{A}_{KKT}$ and $\mathcal{P}_2^{-1}\mathcal{A}_{KKT}$ are included, and this for varying values of the parameter $\gamma$. We have chosen three particular cut-off values for $\gamma$, e.g.

$$\frac{\lambda_{\max}(A)}{100}, \quad \frac{\lambda_{\max}(A)}{1000} \quad \text{and} \quad \frac{\lambda_{\max}(A)}{10000},$$

where $\lambda_{\max}(A) \approx 3.8$, and we indicate the corresponding number $p$ of eigenvalues in the $(1,1)$ block $A$ less than $\gamma$, as well as the condition number of the resulting preconditioned matrices. It is interesting to understand the trade off between the size of $U_\gamma$ (recalling that the columns of $U_\gamma$ are the orthonormal sets of eigenvectors corresponding to the eigenvalues less than $\gamma$), given by $p$, that defines the computational weight for our preconditioners, and the tightness of the intervals on the eigenvalues, both depending on $\gamma$. Indeed, larger values of $\gamma$ imply larger values of $p$ and thus the rank-$p$ update in the approximation of the inverse of $A$,

$$A_\gamma^{-1} = U_\gamma \Lambda_\gamma^{-1} U_\gamma^T + \frac{1}{\alpha} I_n \,,$$

requires more computations, while tightening the bounds on the eigenvalues, the most efficient combination of these two being problem dependent.

We can notice that $\mathcal{P}_1$ and $\mathcal{P}_2$ act differently on the spectrum of the matrix. With $\mathcal{P}_1$, it is mostly the positive lower bound that goes to zero as $\gamma$ goes to zero, the other bounds remaining roughly stable, and this is actually predicted by the result in Theorem 4.1. As opposed to that, with $\mathcal{P}_2$, the negative and positive outer bounds grow with decreasing values of $\gamma$, while the inner bounds stay stable. This is also included in the result given by Theorem 4.3. The main difference, that can also be seen from these two theorems, is that in the case of

| $\gamma$ | $p = \lvert\{\lambda_i \leq \gamma\}\rvert$ | $\mathrm{Spec}(\mathcal{P}_1^{-1}\mathcal{A}_{KKT})$ | $\kappa_2(\mathcal{P}_1^{-1}\mathcal{A}_{KKT})$ |
|---|---|---|---|
| $\frac{\lambda_{\max}(A)}{100}$ | 42 | $[-0.90, -0.45] \cup [9.7\,10^{-2}, 3.28]$ | 33.81 |
| $\frac{\lambda_{\max}(A)}{1000}$ | 33 | $[-0.98, -0.43] \cup [1.1\,10^{-2}, 3.40]$ | 309.09 |
| $\frac{\lambda_{\max}(A)}{10000}$ | 23 | $[-0.99, -0.42] \cup [2.5\,10^{-3}, 3.52]$ | 1408.00 |

| $\gamma$ | $p = \lvert\{\lambda_i \leq \gamma\}\rvert$ | $\mathrm{Spec}(\mathcal{P}_2^{-1}\mathcal{A}_{KKT})$ | $\kappa_2(\mathcal{P}_2^{-1}\mathcal{A}_{KKT})$ |
|---|---|---|---|
| $\frac{\lambda_{\max}(A)}{100}$ | 42 | $[-3.57, -0.40] \cup [1, 4.57]$ | 11.43 |
| $\frac{\lambda_{\max}(A)}{1000}$ | 33 | $[-12.12, -0.39] \cup [1, 13.12]$ | 33.64 |
| $\frac{\lambda_{\max}(A)}{10000}$ | 23 | $[-27.20, -0.38] \cup [1, 28.20]$ | 74.21 |

Table 4.1 – True eigenvalues clustering and condition number of $\mathcal{P}_1^{-1}\mathcal{A}_{KKT}$ and $\mathcal{P}_2^{-1}\mathcal{A}_{KKT}$ for varying values of $\gamma$.

$\mathcal{P}_1$, the inner positive bound is in $\mathcal{O}(\gamma/\alpha)$, whereas with $\mathcal{P}_2$, the outer bounds are in $\mathcal{O}(\sqrt{\alpha/\gamma})$, which grows more slowly than $\mathcal{O}(\sqrt{\gamma/\alpha})$.

Next, the bounds on the intervals in terms of $\lambda_{\max}(A), \lambda_{\min}(A), \gamma$ and $\alpha$ given by Theorem 4.1 and Theorem 4.3 ensure that the eigenvalues of $\mathcal{P}_1^{-1}\mathcal{A}_{KKT}$ and $\mathcal{P}_2^{-1}\mathcal{A}_{KKT}$ belong to

$$[-0.98, -0.28] \cup [0.03, 3.56] \quad \text{and} \quad [-5.13, -0.19] \cup \{1\} \cup [1.19, 6.13]\,,$$

respectively, in the case where $\gamma = \frac{\lambda_{\max}(A)}{100}$. Figure 4.1 shows the eigenvalues distribution of $\mathcal{P}_1^{-1}\mathcal{A}_{KKT}$ and $\mathcal{P}_2^{-1}\mathcal{A}_{KKT}$ with the theoretical bounds, which implies that the condition number of $\mathcal{P}_1^{-1}\mathcal{A}_{KKT}$ is majorated by 119 and that of $\mathcal{P}_2^{-1}\mathcal{A}_{KKT}$ by 32.3.

We now illustrate, in Figure 4.2, the behaviour of preconditioned MINRES on these test cases. Assuming that the eigenvalues of the preconditioned system $\mathcal{P}^{-1}\mathcal{A}_{KKT}$ are bounded within two intervals of the same length, $[-a, -b] \cup [c, d]$ with $a, b, c, d > 0$, we recall from Elman et al. (2005), Section 6.2.4, that the convergence profile of the preconditioned MINRES method is bounded (in exact arithmetic) by

$$\frac{\|r^{2k}\|_{\mathcal{P}^{-1}}}{\|r^0\|_{\mathcal{P}^{-1}}} \leq 2\left(\frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}\right)^k, \tag{4.17}$$

where $\mathcal{P}$ denotes the preconditioning matrix for the KKT system $\mathcal{A}_{KKT}$, and where $r^{2k} = b - \mathcal{A}_{KKT}x^{2k}$ denotes the residual of system (1.3) after $2k$ iterations. The relation (4.17) guarantees that, for a relative residual fixed to $\|r^{2k}\|_{\mathcal{P}^{-1}}/\|r^0\|_{\mathcal{P}^{-1}} = 10^{-q}$, the number of MINRES iteration $k$ is bounded by

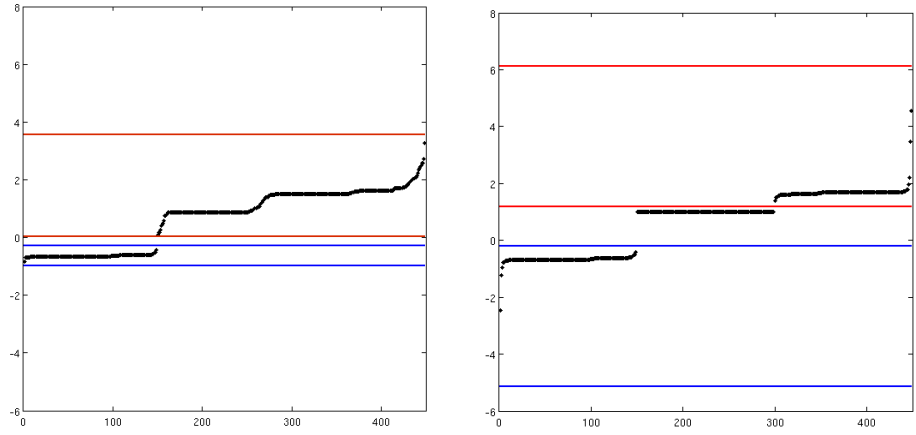


Figure 4.1 – The left-hand subplot shows the eigenvalues distribution of $\mathcal{P}_1^{-1}\mathcal{A}_{KKT}$ and the right-hand subplot the eigenvalues distribution of $\mathcal{P}_2^{-1}\mathcal{A}_{KKT}$.

$$k \quad \leq \quad \frac{-(q+\log_{10} 2)}{\log_{10}\left(\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}\right)}.$$

For comparison purposes, we indicate in Figure 4.2, in dashed lines, the convergence profile corresponding to the upper bound in (4.17) for the various cases. The iterations are stopped when the scaled residual in $\mathcal{P}^{-1}$-norm (with either $\mathcal{P}_1$ or $\mathcal{P}_2$) $\|r^k\|_{\mathcal{P}^{-1}}/\|r^0\|_{\mathcal{P}^{-1}}$ is less than $10^{-8}$. We can observe that smaller values of $\gamma$ increase the number of iterations in both cases. However $\mathcal{P}_2$ seems to be less sensitive than $\mathcal{P}_1$ to small values of $\gamma$, which may be related to the observation above with respect to the behaviour of the bounds in the eigenvalues intervals.

We also mention that, in this particular test case, the scaled residuals in 2-norm $\|r^k\|_2/\|r^0\|_2$ of the unpreconditioned Minres iteration stagnate above $10^{-4}$ without convergence. This can be seen in Figure 4.3, where we also plot, for sake of comparison, the convergence profile of preconditioned Minres with the classical preconditioner (see Golub et al., 2006, example 4.2)

$$\mathcal{P}_{IBB} := \begin{bmatrix} I_n & 0 \\ 0 & B^TB \end{bmatrix}, \tag{4.18}$$

whose purpose is to orthogonalize the constraints. Indeed, note that $\mathcal{P}_{IBB}^{-1}\mathcal{A}_{KKT}$ is similar to $\mathcal{P}_{IBB}^{-1/2}\mathcal{A}_{KKT}\mathcal{P}_{IBB}^{-1/2}$, with

$$\mathcal{P}_{IBB}^{-1/2}\mathcal{A}_{KKT}\mathcal{P}_{IBB}^{-1/2} = \begin{bmatrix} A & B(B^TB)^{-1/2} \\ (B^TB)^{-1/2}B^T & 0 \end{bmatrix},$$

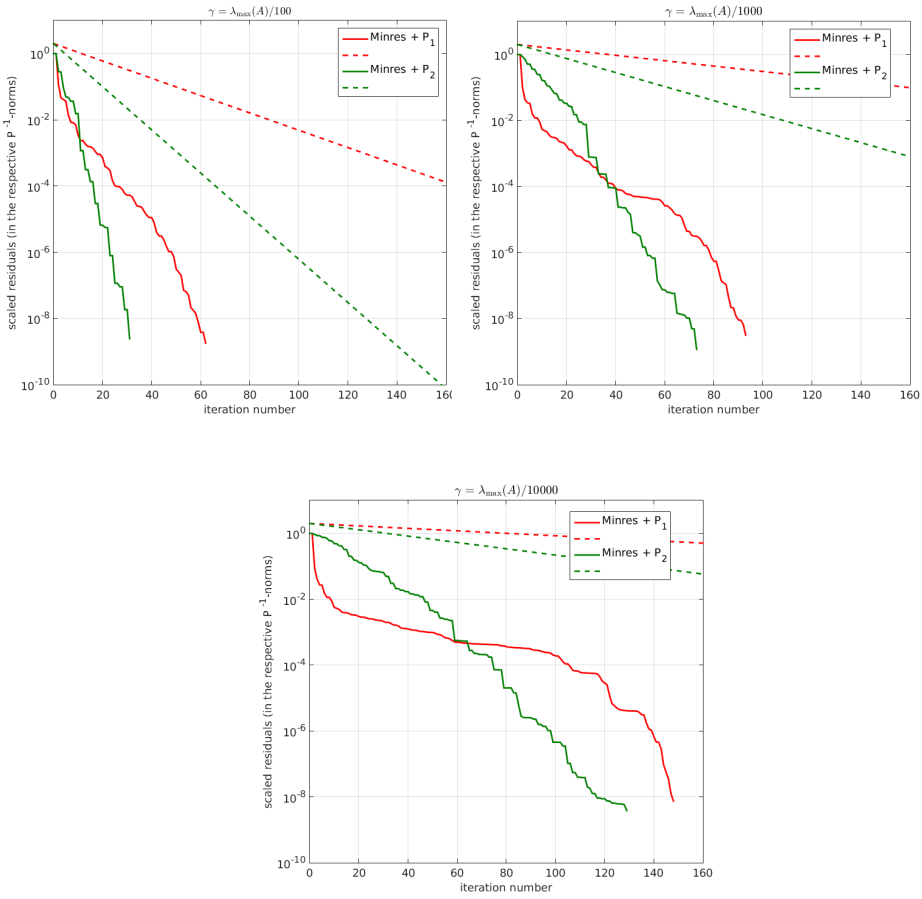



Figure 4.2 – Convergence profiles of preconditioned MINRES with preconditioners $\mathcal{P}_1$ and $\mathcal{P}_2$, for different values of $\gamma$.

where $Q = B(B^TB)^{-1/2}$ satisfies $Q^TQ = I_m$.

The convergence curves for $\mathcal{P}_1$ and $\mathcal{P}_2$ in this figure have been obtained with a value of $\gamma = \lambda_{\max}(A)/100$, and for a fair comparison, the scaled residuals in this figure have been computed in the 2-norm in all cases. We can see that the convergence behaviour of MINRES for $\mathcal{P}_{IBB}^{-1}\mathcal{A}_{KKT}$ is not enough to reduce the iteration number of MINRES. It is then crucial to take account spectral information from $A$ through $A_\gamma^{-1}$ in $\mathcal{P}_1^{-1}$ and $A^{-1}$ in $\mathcal{P}_2^{-1}$, to ensure that the iteration number is small.

Table 4.2 provides the true negative and the positive intervals in which the eigenvalues of $\mathcal{A}_{KKT}$, $\mathcal{P}_1^{-1}\mathcal{A}_{KKT}$, $\mathcal{P}_2^{-1}\mathcal{A}_{KKT}$ and $\mathcal{P}_{IBB}^{-1}\mathcal{A}_{KKT}$ are included. We can see that the ill-conditioning of $\mathcal{P}_{IBB}^{-1}\mathcal{A}_{KKT}$ is caused by the lower bound on

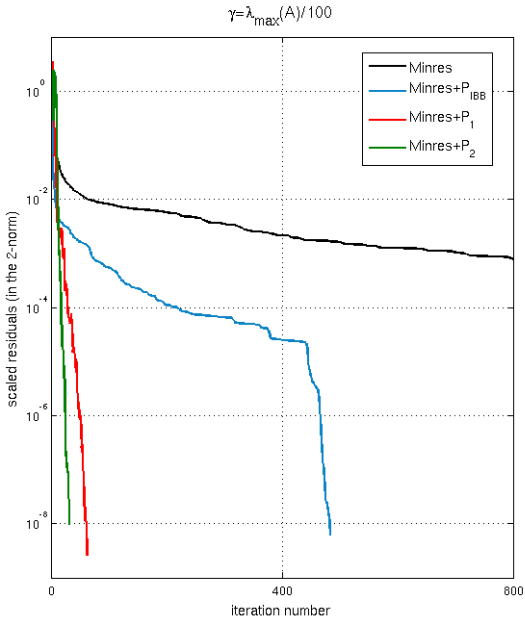



Figure 4.3 – Convergence profiles of MINRES for $\gamma = \lambda_{\max}(A)/100$ (with and without preconditioning).

the positive interval.

| $\mathcal{P}$ | $\mathrm{Spec}(\mathcal{P}^{-1}\mathcal{A}_{KKT})$ | | | $\kappa(\mathcal{P}^{-1}\mathcal{A}_{KKT})$ |
|---|---|---|---|---|
| / | $[-1.6, -2.4\,10^{-6}]$ | $\cup$ | $[2.3\,10^{-5}, 3.8]$ | $1.6\,10^{6}$ |
| $\mathcal{P}_1$ | $[-0.9, -0.5]$ | $\cup$ | $[9.7\,10^{-2}, 3.3]$ | 33.8 |
| $\mathcal{P}_2$ | $[-3.6, -0.4]$ | $\cup$ | $[1.0, 4.6]$ | 11.4 |
| $\mathcal{P}_{IBB}$ | $[-1.0, -0.4]$ | $\cup$ | $[3.0\,10^{-5}, 3.8]$ | $1.3\,10^{5}$ |

Table 4.2 – True eigenvalues clustering and condition number of $\mathcal{A}_{KKT}$, $\mathcal{P}_1^{-1}\mathcal{A}_{KKT}$, $\mathcal{P}_2^{-1}\mathcal{A}_{KKT}$ and $\mathcal{P}_{IBB}^{-1}\mathcal{A}_{KKT}$.

## 4.4 Comparison of the spectral preconditioners for the SQD systems

Now, we illustrate and compare the effectiveness of the two preconditioning alternatives on the previously introduced test example (see Section 3.2.2) for the SQD systems. In the same way, we provide, in Table 4.3, the true negative and positive intervals in which the eigenvalues of $\mathcal{P}_1^{-1}\mathcal{A}_{SQD}$ and $\mathcal{P}_2^{-1}\mathcal{A}_{SQD}$ are included, and this for varying values of the parameter $\gamma$.

| $\gamma$ | $p = \lvert\{\lambda_i \leq \gamma\}\rvert$ | $\mathrm{Spec}(\mathcal{P}_1^{-1}\mathcal{A}_{SQD})$ | $\kappa_2(\mathcal{P}_1^{-1}\mathcal{A}_{SQD})$ |
|---|---|---|---|
| $\frac{\lambda_{\max}(A)}{100}$ | 42 | $[-1.00, -0.49] \cup [9.3\,10^{-2}, 3.28]$ | 35.13 |
| $\frac{\lambda_{\max}(A)}{1000}$ | 33 | $[-1.00, -0.48] \cup [1.0\,10^{-2}, 3.40]$ | 329.87 |
| $\frac{\lambda_{\max}(A)}{10000}$ | 23 | $[-0.99, -0.46] \cup [1.4\,10^{-3}, 3.52]$ | 2598.48 |
| $\gamma$ | $p = \lvert\{\lambda_i \leq \gamma\}\rvert$ | $\mathrm{Spec}(\mathcal{P}_2^{-1}\mathcal{A}_{SQD})$ | $\kappa_2(\mathcal{P}_2^{-1}\mathcal{A}_{SQD})$ |
| $\frac{\lambda_{\max}(A)}{100}$ | 42 | $[-2.90, -0.45] \cup [1, 3.69]$ | 8.16 |
| $\frac{\lambda_{\max}(A)}{1000}$ | 33 | $[-11.05, -0.44] \cup [1, 12.00]$ | 27.17 |
| $\frac{\lambda_{\max}(A)}{10000}$ | 23 | $[-18.04, -0.43] \cup [1, 18.90]$ | 43.85 |

Table 4.3 – True eigenvalues clustering and condition number of $\mathcal{P}_1^{-1}\mathcal{A}_{SQD}$ and $\mathcal{P}_2^{-1}\mathcal{A}_{SQD}$ for varying values of $\gamma$.

Similarly to the KKT systems, we observe that, $\mathcal{P}_1$ and $\mathcal{P}_2$ act differently on the spectrum of the matrix. Indeed, the positive lower bound associated with $\mathcal{P}_1$ goes to zero when $\gamma$ goes to zero, the other bounds remaining roughly stable, and this is actually predicted by the result in Theorem 4.4. The negative and positive outer bounds associated with $\mathcal{P}_2$, grow with decreasing values of $\gamma$, while the inner bounds stay stable. This is also included in the result given by Theorem 4.5. We can also see that the main difference from these two theorems, is that in the case of $\mathcal{P}_1$, the inner positive bound is in $\mathcal{O}(\gamma/\alpha)$, whereas with $\mathcal{P}_2$, the outer bounds are in $\mathcal{O}(\sqrt{\alpha/\gamma})$.

The bounds on the intervals in terms of $\lambda_{\max}(A), \lambda_{\min}(A), \gamma$ and $\alpha$ given by Theorem 4.4 and Theorem 4.5 ensure that the eigenvalues of $\mathcal{P}_1^{-1}\mathcal{A}_{KKT}$ and $\mathcal{P}_2^{-1}\mathcal{A}_{KKT}$ belong to

$$[-1.00, -0.28] \cup [0.03, 3.28] \quad \text{and} \quad \left[-3.23, -1.4\,10^{-4}\right] \cup [1, 3.62]\,,$$

respectively, in the case where $\gamma = \frac{\lambda_{\max}(A)}{100}$. Figure 4.4 shows (using the same scale) the eigenvalues distribution of $\mathcal{P}_1^{-1}\mathcal{A}_{KKT}$ and $\mathcal{P}_2^{-1}\mathcal{A}_{KKT}$ with the theoretical bounds, which implies that the condition number of $\mathcal{P}_1^{-1}\mathcal{A}_{KKT}$ is majo-

rated by 1.1 $10^2$ and this of $\mathcal{P}_2^{-1}\mathcal{A}_{KKT}$ by 2.6 $10^4$. In Figure 4.5, we plot the convergence profile of Minres.

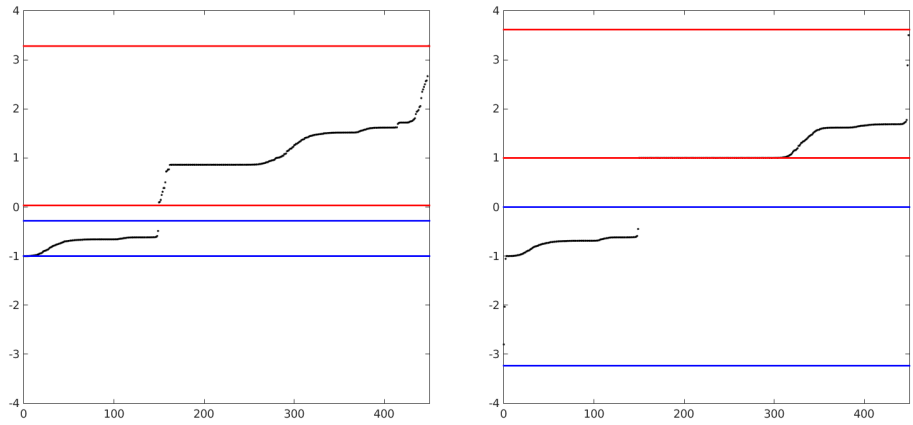


Figure 4.4 – The left-hand subplot shows the eigenvalues distribution of $\mathcal{P}_1^{-1}\mathcal{A}_{SQD}$ and the right-hand subplot the eigenvalues distribution of $\mathcal{P}_2^{-1}\mathcal{A}_{SQD}$.

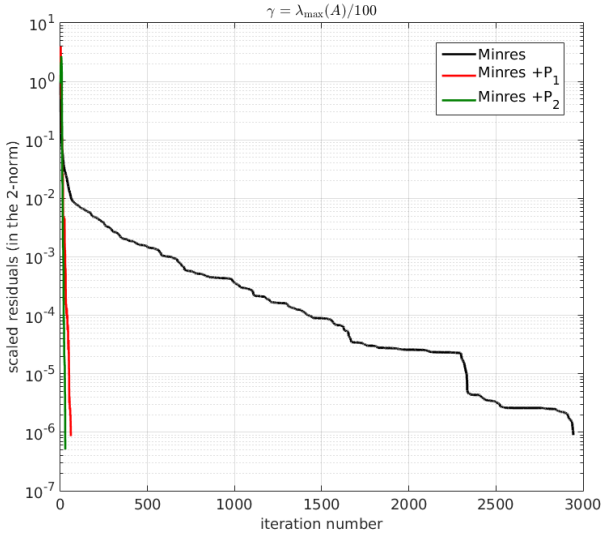



Figure 4.5 – Convergence profiles of Minres for $\gamma = \lambda_{\max}(A)/100$ (with and without preconditioning).

## 4.5 First level preconditioners

In this section, we study the effect of a first level of preconditioning on the matrix $\mathcal{A}_{KKT}$ in (4.2). As mentioned in the previous sections, a first level of preconditioning can be useful to improve the clustering of the eigenvalues of $A$. Considering a block diagonal form for this first level of preconditioning, we denote it by $\mathcal{P}_0$, given as

$$\mathcal{P}_0 := \begin{bmatrix} M & 0 \\ 0 & N \end{bmatrix}, \tag{4.19}$$

in which the two symmetric positive definite matrices $M$ and $N$ are given in a factorized form, $M = R^T R$ and $N = L^T L$, respectively. Let us first rewrite the preconditioned system

$$\mathcal{P}_0^{-1} \mathcal{A}_{KKT} x = \mathcal{P}_0^{-1} b$$

in a symmetrized manner as

$$\hat{\mathcal{A}}_{KKT} \hat{x} = \hat{b} \equiv \begin{bmatrix} \hat{A} & \hat{B} \\ \hat{B}^T & 0 \end{bmatrix} \begin{bmatrix} Ru \\ Lv \end{bmatrix} = \begin{bmatrix} R^{-T} f \\ L^{-T} g \end{bmatrix} \tag{4.20}$$

with $\hat{A} = R^{-T} A R^{-1}$ being the symmetric positive definite matrix corresponding to $A$ preconditioned with $M$, and $\hat{B} = R^{-T} B L^{-1}$ being the corresponding preconditioned constraint matrix. In this case, the associated Schur complement becomes $\hat{S} = \hat{B}^T \hat{A}^{-1} \hat{B}$.

In the following section, we focus on the KKT systems and we derive the formulations of $A_\gamma^{-1}$ and $S_\gamma^{-1}$, the approximations of the inverse of $\hat{A}$ and of the Schur complement associated to the system (4.20) as introduced in Sections 3.1 and 3.2, respectively.

### 4.5.1 Combination of a first level preconditioner with spectral approximations

Similarly to the eigendecomposition of the matrix $A$ in (3.3), we split the spectrum of $\hat{A}$ in two parts, with $\hat{\Lambda}_\gamma \in \mathbb{R}^{p \times p}$ the diagonal matrix containing the $p$ eigenvalues less than a given positive number $\gamma \in [\lambda_{\min}(\hat{A}), \lambda_{\max}(\hat{A})]$. We assume that the first level of preconditioning ensures that the number $p$ of these eigenvalues is small. Following the steps in Section 3.1, we approximate the inverse of $\hat{A}$ as in (3.4) with

$$\hat{A}_\gamma^{-1} \quad = \quad \hat{U}_\gamma \hat{\Lambda}_\gamma^{-1} \hat{U}_\gamma^T + \frac{1}{\alpha} I_n, \tag{4.21}$$

where $\hat{U}_\gamma$ denotes the set of eigenvectors of $\hat{A}$ associated to those eigenvalues below the given threshold $\gamma$ and $\alpha > 0$ is some estimate of the average of the remaining eigenvalues (those not in $\hat{\Lambda}_\gamma$), or of $\lambda_{\max}(\hat{A})$. Similarly to Section 3.2.1,

we next derive the approximation of the inverse of the Schur complement $\hat{S}$ as in (3.13) with

$$\hat{S}_\gamma^{-1} = \alpha \left(\hat{B}^T \hat{B}\right)^{-1/2} \left( I_m - \hat{K}_\gamma \left( \frac{1}{\alpha} \hat{\Lambda}_\gamma + \hat{K}_\gamma^T \hat{K}_\gamma \right)^{-1} \hat{K}_\gamma^T \right) \left(\hat{B}^T \hat{B}\right)^{-1/2} \quad (4.22)$$

where $\hat{K}_\gamma$ is the operator defined by $(\hat{B}^T \hat{B})^{-1/2} \hat{B}^T \hat{U}_\gamma$. Note that the eigenvalue equation $\hat{A}\hat{U}_\gamma = \hat{U}_\gamma \hat{\Lambda}_\gamma$ together with $\hat{A} = R^{-T} A R^{-1}$ can be written as

$$R^{-T} A R^{-1} \hat{U}_\gamma = \hat{U}_\gamma \hat{\Lambda}_\gamma,$$

which is equivalent to

$$A R^{-1} \hat{U}_\gamma = R^T \hat{U}_\gamma \hat{\Lambda}_\gamma.$$

Introducing the notation $V_\gamma = R^{-1}\hat{U}_\gamma$ , we have the generalized eigenvalue equation

$$A V_\gamma = M V_\gamma \hat{\Lambda}_\gamma. \quad (4.23)$$

We deduce that the matrix $V_\gamma$ corresponds to the eigenvectors of $A$ preconditioned with $M$. We can rewrite (4.22) as

$$\hat{S}_\gamma^{-1} = \alpha \left( \left(\hat{B}^T \hat{B}\right)^{-1} - \left(\hat{B}^T \hat{B}\right)^{-1/2} \hat{K}_\gamma \left( \frac{1}{\alpha} \hat{\Lambda}_\gamma + \hat{K}_\gamma^T \hat{K}_\gamma \right)^{-1} \hat{K}_\gamma^T \left(\hat{B}^T \hat{B}\right)^{-1/2} \right),$$

with

$$\begin{aligned} \hat{B}^T \hat{B} &= L^{-T} B^T R^{-1} R^{-T} B L^{-1} \\ &= L^{-T} \left( B^T M^{-1} B \right) L^{-1} \\ &= L^{-T} S_M L^{-1}, \end{aligned}$$

where we use the notation $S_M = B^T M^{-1} B$. Using the fact that

$$\begin{aligned} (\hat{B}^T \hat{B})^{-1/2} \hat{K}_\gamma &= (\hat{B}^T \hat{B})^{-1} \hat{B}^T \hat{U}_\gamma \\ &= (\hat{B}^T \hat{B})^{-1} L^{-T} B^T R^{-1} R V_\gamma \\ &= L S_M^{-1} B^T V_\gamma, \end{aligned}$$

and

$$\begin{aligned} \hat{K}_\gamma^T \hat{K}_\gamma &= \hat{U}_\gamma^T \hat{B} (\hat{B}^T \hat{B})^{-1} \hat{B}^T \hat{U}_\gamma \\ &= V_\gamma^T R^T R^{-T} B L^{-1} (\hat{B}^T \hat{B})^{-1} L^{-T} B^T R^{-1} R V_\gamma \\ &= V_\gamma^T B S_M^{-1} B^T V_\gamma, \end{aligned}$$

$\hat{S}_\gamma^{-1}$ becomes,

$$\hat{S}_\gamma^{-1} \;=\; \alpha\left(LS_M^{-1}L^T - LS_M^{-1}B^TV_\gamma\left(\frac{1}{\alpha}\hat{\Lambda}_\gamma + V_\gamma^TBS_M^{-1}B^TV_\gamma\right)^{-1}V_\gamma^TBS_M^{-1}L^T\right).$$

We can then introduce the operator $Z_\gamma$ given by

$$Z_\gamma \;=\; S_M^{-1/2}B^TV_\gamma = (B^TM^{-1}B)^{-1/2}B^TV_\gamma, \tag{4.24}$$

which plays the role of $K_\gamma$ with respect to the eigenvectors of $A$ preconditioned with $M$ and to the modified inner product associated to the first level of preconditioning $M$, to write

$$\hat{S}_\gamma^{-1} \;=\; \alpha LS_M^{-1/2}\left(I_m - Z_\gamma\left(\frac{1}{\alpha}\hat{\Lambda}_\gamma + Z_\gamma^TZ_\gamma\right)^{-1}Z_\gamma^T\right)S_M^{-1/2}L^T. \tag{4.25}$$

In the next section, we use the approximations $\hat{A}_\gamma^{-1}$ and $\hat{S}_\gamma^{-1}$ to construct new approximations of block diagonal preconditioners.

### 4.5.2 Combination of a first level preconditioner with spectral preconditioners

In this section, we show that the combination of the two levels of preconditioning,

$$\mathcal{P}_i^{-1}\mathcal{P}_0^{-1}\mathcal{A}u = \mathcal{P}_i^{-1}\mathcal{P}_0^{-1}b, \tag{4.26}$$

with $i$ either 1 or 2, can be expressed in a simple form

$$\mathcal{P}_{iM}^{-1}\mathcal{A}x = \mathcal{P}_{iM}^{-1}b,$$

with $\mathcal{P}_{iM}$ a block diagonal preconditioner satisfying $\mathcal{P}_{iM}^{-1} = \mathcal{P}_i^{-1}\mathcal{P}_0^{-1}$. Using the approximation of the inverse of $\hat{A}$ and the Schur complement $\hat{S}_\gamma$ developped in Section 4.5, we apply either

$$\hat{\mathcal{P}}_1^{-1} = \begin{bmatrix} \hat{A}_\gamma^{-1} & \\ & \hat{S}_\gamma^{-1} \end{bmatrix} \quad\text{and}\quad \hat{\mathcal{P}}_2^{-1} = \begin{bmatrix} \hat{A}^{-1} & \\ & \hat{S}_\gamma^{-1} \end{bmatrix}$$

with $\hat{A}_\gamma^{-1}$ and $\hat{S}_\gamma^{-1}$ defined by (4.21) and (4.25) respectively, to the symmetrized system $\hat{\mathcal{A}}\hat{u} = \hat{b}$. It is then easy to see that the matrix $L$ cancels with its inverse or that $N = LL^T$ can be eliminated from both sides of the system of equations, and therefore the combined two levels of preconditioning can be condensed into the following formulations

$$\mathcal{P}_{1M}^{-1}\mathcal{A}u = \mathcal{P}_{1M}^{-1}b \quad\text{and}\quad \mathcal{P}_{2M}^{-1}\mathcal{A}u = \mathcal{P}_{2M}^{-1}b$$

where

$$\mathcal{P}_{1M}^{-1} = \begin{bmatrix} V_\gamma \hat{\Lambda}_\gamma^{-1} V_\gamma^T + \frac{1}{\alpha} M^{-1} & 0 \\ 0 & \alpha S_M^{-1/2} \left( I_m - Z_\gamma \left( \frac{1}{\alpha} \hat{\Lambda}_\gamma + Z_\gamma^T Z_\gamma \right)^{-1} Z_\gamma^T \right) S_M^{-1/2} \end{bmatrix}$$

and

$$\mathcal{P}_{2M}^{-1} = \begin{bmatrix} A^{-1} & 0 \\ 0 & \alpha S_M^{-1/2} \left( I_m - Z_\gamma \left( \frac{1}{\alpha} \hat{\Lambda}_\gamma + Z_\gamma^T Z_\gamma \right)^{-1} Z_\gamma^T \right) S_M^{-1/2} \end{bmatrix}.$$

We first observe that the inverses of $\mathcal{P}_{1M}$ and $\mathcal{P}_{2M}$ do not depend on the first level of preconditioning $N$ on $B$, implying that the preconditioning of $B$ is not necessary (except perhaps for numerical issues, like scaling, etc.). Obviously, the bad conditioning of $B$ is taken into account in the $S_M^{-1/2} = \left(B^T M^{-1} B\right)^{-1/2}$ terms that appear on both sides of the $(2,2)$ block in $\mathcal{P}_{1M}^{-1}$ and $\mathcal{P}_{2M}^{-1}$. Finally, we can easily get rid of the matrices square roots in this $(2,2)$ block and rewrite it as

$$\alpha \left( S_M^{-1} - S_M^{-1} (B^T U_\gamma) \left( \frac{1}{\alpha} \hat{\Lambda}_\gamma + (B^T U_\gamma)^T S_M^{-1} (B^T U_\gamma) \right)^{-1} (B^T U_\gamma)^T S_M^{-1} \right), \quad (4.27)$$

where the major computational part is in fact to solve linear systems with matrix

$$S_M = B^T M^{-1} B$$

in an appropriate manner.

Concerning the $(1,1)$ block of $\mathcal{P}_{1M}^{-1}$, we recall that $p \ll n$ so that it resumes in a low rank update to $\frac{1}{\alpha} M^{-1}$. We also note that either $M$ is available in a factorized form or the solution with $M$ is easy in principle. A way to extract a good approximation for the partial spectral information $\hat{\Lambda}_\gamma$ and $V_\gamma$ within Krylov techniques is proposed in Golub et al. (2007). With respect to the $(1,1)$ block of $\mathcal{P}_{2M}^{-1}$, this partial spectral information can be used to efficiently solve with matrix $A$ in preconditioned or deflated Krylov techniques (see for instance Giraud et al., 2006).

Coming back to the solution with matrix $S_M = B^T M^{-1} B$, a lot of attention has been devoted to this issue with, in particular, the application of constraint preconditioners of the form

$$\mathcal{P}_c = \begin{bmatrix} M & B \\ B^T & 0 \end{bmatrix}, \quad (4.28)$$

and we refer to Benzi et al. (2005) for a nice survey. The nice feature of the Schur complement approximation in (4.27) is that the very small low rank update added to $S_M^{-1}$ incorporates the missing information whenever the first level of preconditioning is not enough to set up Minres in good conditions

for linear convergence. With respect to computational costs, this low rank update can be constructed once, factorized beforehand, and reused many times to speed up solutions with changing right-hand sides and the same coefficient matrix. In this particular context, the extra cost to build the various spectral components required in this approach can be very rapidly amortized. This has already been illustrated in the case of ill-conditioned symmetric positive definite systems (see Golub et al., 2007).

Chapter 5

# Stokes problem

In this chapter, we consider a problem in fluid dynamics generated by the `Matlab` package `ifiss` produced by Elman, Ramage and Silvester (2002) written jointly with the book of Elman et al. (2005). The toolbox models a steady incompressible fluid flow using partial differential equations (PDEs). It includes algorithms for discretization by finite element methods, which are used to rewrite the problem as a linear system of equations. We use the well-know Stokes problem to illustrate the numerical behaviour of the spectral preconditioners introduced in Chapter 4. We consider a viscous fluid moving in a domain of the two dimensional space $\Omega$. The Stokes problem is governed by PDEs given by

$$\begin{aligned} -\nabla^2 \vec{u} + \nabla p &= \vec{0}, & (5.1)\\ \nabla \cdot \vec{u} &= 0, & (5.2) \end{aligned}$$

with boundary conditions

$$\begin{aligned} \vec{u} &= \vec{w} \quad \text{on} \quad \partial\Omega_D,\\ \frac{\partial \vec{u}}{\partial n} - \vec{n}p &= \vec{0} \quad \text{on} \quad \partial\Omega_N, \end{aligned}$$

where $\partial\Omega_D \cup \partial\Omega_N = \partial\Omega$ and $\partial\Omega_D$ and $\partial\Omega_N$ are distinct. The vector valued function $\vec{u}$ represents the fluid velocity and the scalar function $p$ represents the pressure. Equation (5.1) is the conservation of the momentum of the fluid, while the second equation (5.2) enforces conservation of mass (also referred to as the incompressibility constraint). This last equation characterizes the "low-speed" flow as for instance engine oil. In the `Ifiss` software, a finite element discretization is used to express the Stokes problem as a system of linear equations with the following KKT form,

$$\begin{bmatrix} A & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix},$$

where $A \in \mathbb{R}^{n_u \times n_u}$ is called the *vector Laplacian matrix* and $B \in \mathbb{R}^{n_u \times n_p}$ is called the *divergence matrix*.

Different discretization approaches are discussed in Elman et al. (2005), Section 5.3. In this chapter, we first consider the ($Q_2$-$Q_1$) Taylor-Hood method and next the $Q_1$-$P_0$ method. In this last case, the element pair is unstable. In such cases, one can use stabilization techniques leading to Stokes problem with the following SQD form

$$\begin{bmatrix} A & B \\ B^T & -\beta C \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}, \tag{5.3}$$

where $\beta > 0$ is a stabilization parameter.

In Section 5.1, we analyse and compare the behaviour of preconditioned MINRES by $\mathcal{P}_1$ and $\mathcal{P}_2$ introduced in Chapter 4 for KKT systems. Section 5.2 illustrates $\mathcal{P}_1$ and $\mathcal{P}_2$ on Stokes problems with SQD form.

## 5.1 Preconditioned approach for KKT systems

The KKT system coming from Stokes problem is generated using the `ifiss` package. We present numerical results for a simple test problem arising in incompressible fluid flow: Flow over a backward facing step in L-shaped domain represented in Figure 5.1. The computations were performed on a workstation using `Matlab` *R*2015*a* and we have used `ifiss` 3.4. We consider the Stokes problem as described above, employing the Taylor-Hood elements on a non-uniform grid with grid stretch factor equal to 2. The size of the system is 7235 and we solve the resulting linear system with MINRES. The matrix $A$ is symmetric positive definite of order $n = 6402$ and the rectangular matrix $B$ has size $6402 \times 833$.

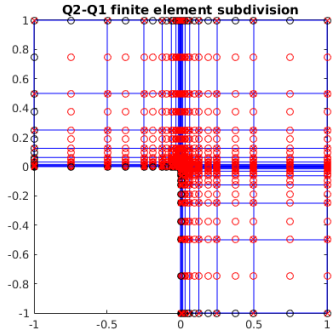



Figure 5.1 – L-shaped domain with non-uniform grid.

As first-level of preconditioning, we consider a Jacobi scaling of $A$ to set the diagonal of the preconditioned matrix to 1. The spectrum of the resulting preconditioned matrix is well clustered, with 246 eigenvalues less than $\gamma = \frac{\lambda_{\max}(A)}{100} \approx 2.7\,10^{-2}$, and with extreme eigenvalues of $1.5\,10^{-6}$ and 32.7. The condition number of $A$ is $2.2\ 10^{7}$. Figure 5.2 shows (on logarithmic scale) the eigenvalues of this preconditioned matrix. For simplicity, we shall denote as $A$ this preconditioned matrix in the following.

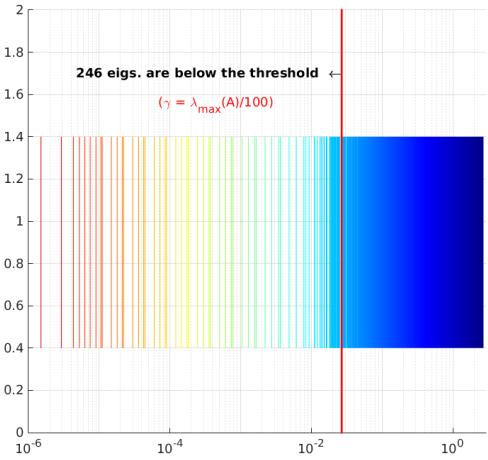



Figure 5.2 – Spectrum of the matrix $A$ generated by Matlab with ifiss package after Jacobi scaling.

We now illustrate, in Figure 5.3, the behaviour of Minres preconditioned by $\mathcal{P}_1$ and $\mathcal{P}_2$ introduced in Chapter 4 using exact spectral information of $A$. Similarly to Section 4.3, we indicate, in dashed lines, the convergence profile corresponding to the upper bound of scaled residuals in the respective $\mathcal{P}^{-1}$-norms (with either $\mathcal{P}_1$ or $\mathcal{P}_2$). The convergence curves for $\mathcal{P}_1$ and $\mathcal{P}_2$ have been obtained with a value of $\gamma = \lambda_{\max}(A)/100$ and the iterations are stopped when the scaled residual in $\mathcal{P}^{-1}$-norm (with either $\mathcal{P}_1$ or $\mathcal{P}_2$) is less than $10^{-8}$. We can observe that $\mathcal{P}_2$ is better than $\mathcal{P}_1$, which may be related as we have seen to the behaviour of the bounds on the eigenvalue intervals given by Theorem 4.1 and Theorem 4.3.

In Figure 5.4, we also plot, for sake of comparison, the convergence profile of preconditioned Minres with the classical preconditioner $\mathcal{P}_{IBB}$ defined by (4.18) in Chapter 4 and the least-squares commutator (LSC) preconditioner (see Elman et al., 2005, Section 8.2.2) defined by

$$\mathcal{P}_{LSC}^{-1} := \begin{bmatrix} A^{-1} & 0 \\ 0 & S_{LSC}^{-1} \end{bmatrix}, \tag{5.4}$$

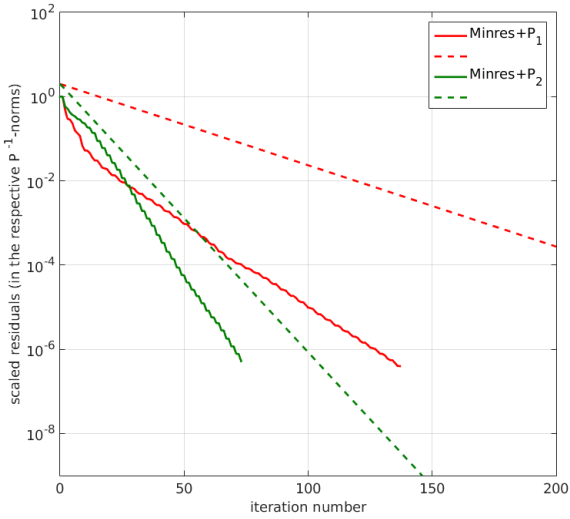



Figure 5.3 – Convergence profiles of MINRES preconditioned with $\mathcal{P}_1$ and $\mathcal{P}_2$ for the Stokes problem of the KKT form.

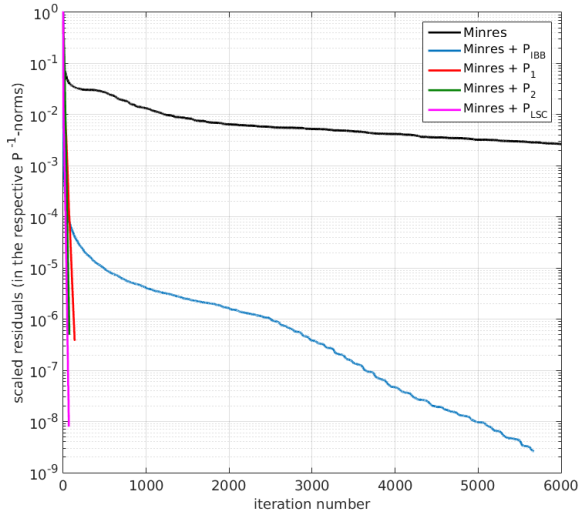



Figure 5.4 – Convergence profiles of MINRES (with and without preconditioning).

where $S_{LSC}^{-1} = (B^TB)^{-1}(B^TAB)(B^TB)^{-1}$ in the unscaled version (which is one of the preconditioning possibilities in `ifiss`). We point out some analogies of the matrix $S_\gamma^{-1}$ introduced in (3.13) with the LSC preconditioner with respect to the ingredients that specifically concern the constraint matrix $B$. Indeed, in both cases the inverse of the Schur complement includes the inverse of $B^TB$ at both ends, but the LSC preconditioner incorporates directly matrix $B^TAB$ instead of the rank-$k$ update that we have proposed to approximate the inverve of the Schur complement.

Considering preconditioners $\mathcal{P}_1$, $\mathcal{P}_2$ and $\mathcal{P}_{LSC}$, Table 5.1 provides the number of iterations of preconditioned Minres. We can see that the preconditioner $\mathcal{P}_2$, including spectral approximation of the inverse of the Schur complement, is close to the number of iterations of $\mathcal{P}_{LSC}$, while $\mathcal{P}_1$, using an approximation of the inverse of $A$, demands twice as many iterations.

| $\mathcal{P}$ | # iter. Minres |
|---|---|
| $\mathcal{P}_1$ | 137 |
| $\mathcal{P}_2$ | 73 |
| $\mathcal{P}_{LSC}$ | 67 |

Table 5.1 – Number of iterations of preconditioned Minres by $\mathcal{P}_1$, $\mathcal{P}_2$ and $\mathcal{P}_{LSC}$.

Table 5.2 provides the true negative and positive intervals in which the eigenvalues of $\mathcal{A}_{KKT}$, $\mathcal{P}_1^{-1}\mathcal{A}_{KKT}$, $\mathcal{P}_2^{-1}\mathcal{A}_{KKT}$, $\mathcal{P}_{IBB}^{-1}\mathcal{A}_{KKT}$ and $\mathcal{P}_{LSC}^{-1}\mathcal{A}_{KKT}$ are included. We can see that the ill-conditioning of $\mathcal{P}_{IBB}^{-1}\mathcal{A}_{KKT}$ is caused by the lower bound on the positive interval. We observe that the condition number of $\mathcal{P}_{LSC}^{-1}\mathcal{A}_{KKT}$ is larger than for $\mathcal{P}_1$ and $\mathcal{P}_2$. The lower bound of positive interval of $\mathcal{P}_1^{-1}\mathcal{A}_{KKT}$ could explain that the number of iterations of $\mathcal{P}_1$ is larger than for $\mathcal{P}_{LSC}$.

| $\mathcal{P}$ | $\mathrm{Spec}(\mathcal{P}^{-1}\mathcal{A}_{KKT})$ | | | $\kappa(\mathcal{P}^{-1}\mathcal{A}_{KKT})$ |
|---|---|---|---|---|
| / | $[-5.6\,10^{-2}, -2.8\,10^{-11}]$ | $\cup$ | $[4.3\,10^{-6}, 2.8]$ | $9.6\,10^{10}$ |
| $\mathcal{P}_1$ | $[-1.0, -0.6]$ | $\cup$ | $[2.7\,10^{-2}, 2.6]$ | $9.4\,10^{1}$ |
| $\mathcal{P}_2$ | $[-4.4, -0.6]$ | $\cup$ | $[1.0, 5.4]$ | $9.4$ |
| $\mathcal{P}_{IBB}$ | $[-3.3, -5.7]$ | $\cup$ | $[3.0\,10^{-5}, 3.8]$ | $1.3\,10^{5}$ |
| $\mathcal{P}_{LSC}$ | $[-2.3\,10^{2}, -0.6]$ | $\cup$ | $[1.0, 2.3\,10^{2}]$ | $3.8\,10^{2}$ |

Table 5.2 – True eigenvalues clustering and condition number of $\mathcal{P}_1^{-1}\mathcal{A}_{KKT}$, $\mathcal{P}_2^{-1}\mathcal{A}_{KKT}$, $\mathcal{P}_{IBB}^{-1}\mathcal{A}_{KKT}$ and $\mathcal{P}_{LSC}^{-1}\mathcal{A}_{KKT}$.

## 5.2 Preconditioned approach for the SQD systems

Now, we illustrate the behaviour of preconditioners $\mathcal{P}_1$ and $\mathcal{P}_2$ on the previous test example with a $Q_1$-$P_0$ discretization leading to the SQD system (5.3). We also consider a Jacobi scaling of $A$ and the spectrum of the resulting preconditioned matrix is well clustered, with 216 eigenvalues less than $\gamma = \frac{\lambda_{\max}}{100} \approx 3\,10^{-2}$, and with extreme eigenvalues of $2.2\,10^{-6}$ and 3. The condition number of $A$ is $1.4\ 10^6$. Figure 5.5 shows (on logarithmic scale) the eigenvalues of this preconditioned matrix.

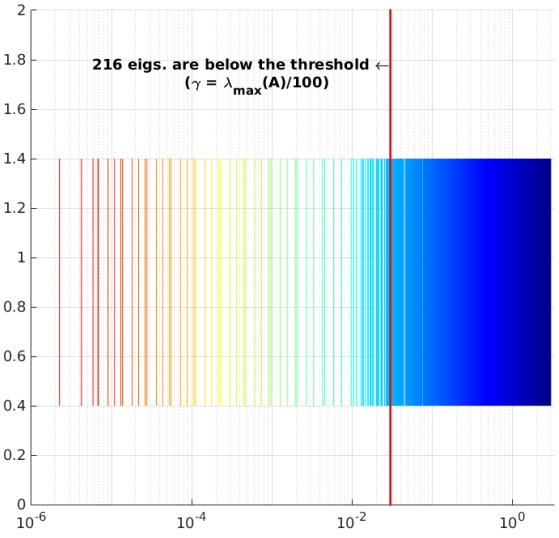



Figure 5.5 – Spectrum of the matrix $A$ generated by Matlab with ifiss package after Jacobi scaling.

In the same way, we provide, in Figure 5.6, the behaviour of MINRES preconditioned by $\mathcal{P}_1$ and $\mathcal{P}_2$. The convergence curves for $\mathcal{P}_1$ and $\mathcal{P}_2$ have been obtained with a value of $\gamma = \lambda_{\max}(A)/100$ and the iterations are stopped when the scaled residual in $\mathcal{P}^{-1}$-norm (with either $\mathcal{P}_1$ or $\mathcal{P}_2$) is less than $10^{-8}$. Similarly to the KKT system, we can observe that $\mathcal{P}_2$ is better than $\mathcal{P}_1$, which may be related as we have seen to the behaviour of the bounds on the eigenvalue intervals given by Theorems 4.4 and 4.5.

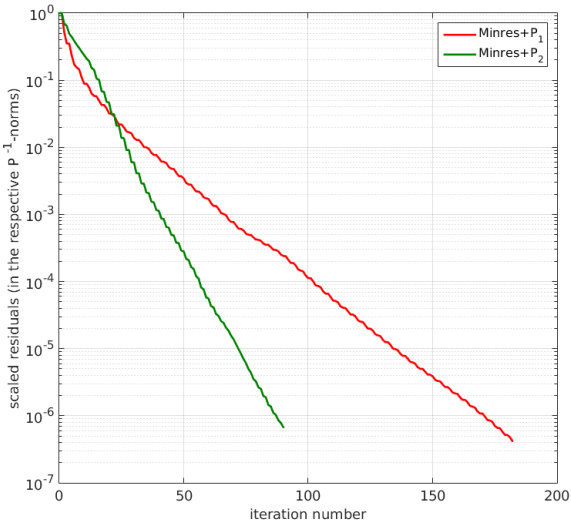



Figure 5.6 – Convergence profiles of Minres preconditioned with preconditioners $\mathcal{P}_1$ and $\mathcal{P}_2$ for the Stokes problem of the SQD form.

Chapter 6

# Interaction between the blocks in KKT matrices

In the previous chapter, we presented two block diagonal preconditioners for KKT systems of matrix

$$\mathcal{A}_{KKT} = \begin{bmatrix} A & B \\ B^T & 0 \end{bmatrix}, \tag{6.1}$$

which incorporate the ill-conditioned part of the matrix $A$ through the approximation of the inverse of the Schur complement $S = B^T A^{-1} B$ and of the inverse of $A$. This proposed approach has the advantage of explicitly showing the special link between the matrices $A$ and $B$. To our knowledge, no study of the interactions between $A$ and $B$ in (6.1) using the approximation of the inverse of the Schur complement introduced in Chapter 3 has been proposed so far. Yet the optimization and linear algebra communities using preconditioning for KKT matrices have knowledge of the interaction between these two matrices. Maybe is this due to numerical experimentations.

The purpose of this chapter is to highlight, from a theoretical point of view, some aspects of the interaction between $A$ and $B$ when solving systems of the form $\mathcal{A}_{KKT} x = b$, as in (1.3). Indeed, it is commonly observed that despite their possible ill-conditioning, some recombination of $A$ and $B$ occurs that sometimes degrades but can also improve the convergence of Krylov subspace methods like MINRES.

Section 6.1 gives a first insight on the interaction between $A$ and $B$ through the Schur complement approximation and shows some configurations according to which the influence of the small eigenvalues of $A$ can have an effect on the convergence of MINRES. The next sections give some intuition on the interaction between the matrices $A$ and $B$ through a toy example first, then for varying constraint matrices. Section 6.3 refines the bounds on the eigenvalues

of $\mathcal{A}_{KKT}$ through new theoretical developments . Finally, the last section of the chapter studies the possibility of reducing the low rank update in the inverse of the approximation of the Schur complement (3.13) and we generalize the block diagonal preconditioner $\mathcal{P}_1$ defined in Chapter 4 in this context.

## 6.1 Interaction between blocks in the Schur complement approximation

We first clarify through the Schur complement approximation $S_\gamma$ introduced in Chapter 3, Section 3.2.1, the link between the matrices $A$ and $B$. We thus come back to the inverse of the approximation of the Schur complement (3.13) given by

$$S_\gamma^{-1} = \alpha(B^TB)^{-1/2}\left(I_m - K_\gamma\left(\frac{1}{\alpha}\Lambda_\gamma + K_\gamma^TK_\gamma\right)^{-1}K_\gamma^T\right)(B^TB)^{-1/2}, \quad (6.2)$$

where $K_\gamma \in \mathbb{R}^{m\times p}$ is the operator defined in (3.14) by $(B^TB)^{-1/2}B^TU_\gamma$ with the constraint matrix $B \in \mathbb{R}^{n\times m}$ and with $U_\gamma \in \mathbb{R}^{n\times p}$, which contains the orthonormal set of the $p$ eigenvectors associated to the eigenvalues in $A$ below a given threshold $\gamma$. As we have seen in Chapter 3, the singular values of $K_\gamma$ correspond to the cosines of the principal angles between the two subspaces $\mathcal{I}m(B)$ and $\mathcal{I}m(U_\gamma)$, since $B(B^TB)^{-1/2}$ represents an orthonormal basis for $\mathcal{I}m(B)$ (see, e.g., Golub and Van Loan, 2013, Section 6.4.3). The expression

$$\frac{1}{\alpha}\Lambda_\gamma + K_\gamma^TK_\gamma$$

in (6.2) explicitly shows the interaction between $A$ and $B$, with the combined effects of both the smallest eigenvalues of $A$ (through $\Lambda_\gamma$) and the cosines of the principal angles between $\mathcal{I}m(B)$ and $\mathcal{I}m(U_\gamma)$ (through $K_\gamma$).

Let us now consider the matrix $K \in \mathbb{R}^{m\times n}$ defined as

$$K = Q^TU = [Q^TU_\gamma,\ Q^T\tilde{U}_\gamma] = [K_\gamma,\ \tilde{K}_\gamma], \quad (6.3)$$

where

$$Q = B(B^TB)^{-1/2} \in \mathbb{R}^{n\times m} \quad (6.4)$$

satisfies $Q^TQ = I_m$ by definition, $U = [U_\gamma,\ \tilde{U}_\gamma]$ is the orthogonal matrix of the eigendecomposition (3.3) of $A$, $K_\gamma$ is the operator used in (6.2) and we set $\tilde{K}_\gamma = Q^T\tilde{U}_\gamma$. The columns of $K^T$ are orthonormal, implying that $K_\gamma K_\gamma^T + \tilde{K}_\gamma\tilde{K}_\gamma^T = I_m$. If we next complete the matrix $K^T$ by $m-n$ orthonormal columns to provide an orthogonal matrix of $\mathbb{R}^{n\times n}$, and if we apply the CS decomposition as in Appendix A or Paige and Saunders (1981), Section 4, one can guarantee

the existence of orthogonal matrices $V_\gamma \in \mathbb{R}^{p\times p}$, $\tilde{V}_\gamma \in \mathbb{R}^{(n-p)\times(n-p)}$ and $W \in \mathbb{R}^{m\times m}$ such that

$$V_\gamma^T K_\gamma^T W = \mathcal{C} = \mathrm{diag}(c_1, \ldots, c_r) \in \mathbb{R}^{p\times m}, \qquad r = \min\{p, m\}, \tag{6.5}$$

and

$$\tilde{V}_\gamma^T \tilde{K}_\gamma^T W = \mathcal{S} = \mathrm{diag}(s_1, \ldots, s_q) \in \mathbb{R}^{(n-p)\times m}, \qquad q = \min\{n-p, m\}, \tag{6.6}$$

where

$$\mathcal{C}^T\mathcal{C} + \mathcal{S}^T\mathcal{S} = I_m. \tag{6.7}$$

The singular values $c_i$ and $s_i$ of $K_\gamma^T$ and $\tilde{K}_\gamma^T$, respectively, are cosines and sines satisfying (without loss of generality)

$$1 \geq c_1 \geq \ldots \geq c_r \geq 0 \qquad \text{and} \qquad 0 \leq s_1 \leq \ldots \leq s_q \leq 1.$$

Among these values, $\min\{r, q\}$ correspond to the cosines and sines of the principal angles between $\mathcal{I}m(B)$ and $\mathcal{I}m(U_\gamma)$, the other values being equal to either 0 or 1, depending on the dimensions $p, m$ and $n$. The associated $\min\{r, q\}$ principal vectors (see Appendix A or Golub and Van Loan, 2013, Section 6.4.3) are defined by the $\min\{r, q\}$ first columns of matrix $U_\gamma V_\gamma$ and matrix $QW$, in $\mathcal{I}m(U_\gamma)$ and $\mathcal{I}m(B)$ respectively. Equation (6.5) now implies that $K_\gamma^T = V_\gamma \mathcal{C} W^T$, and from the expression (6.2) of $S_\gamma^{-1}$ we have that

$$S_\gamma^{-1} = \alpha (B^T B)^{-1/2}\, P\, (B^T B)^{-1/2},$$

where

$$\begin{aligned} P &= I_m - W\mathcal{C}^T V_\gamma^T \left(\frac{1}{\alpha}\Lambda_\gamma + V_\gamma \mathcal{C}\mathcal{C}^T V_\gamma^T\right)^{-1} V_\gamma \mathcal{C} W^T \\ &= I_m - W\mathcal{C}^T \left(\frac{1}{\alpha} V_\gamma^T \Lambda_\gamma V_\gamma + \mathcal{C}\mathcal{C}^T\right)^{-1} \mathcal{C} W^T. \end{aligned}$$

At this point, several configurations can occur, depending on whether the size $p$ of the invariant subspace $U_\gamma$ is smaller than the number $m$ of constraint equations or not, and/or whether some cosines are zero, meaning that there exists some orthogonality between the principal vectors themselves. In this last case or when $p > m$, the result is that some rows in the $p \times m$ rectangular matrix $\mathcal{C}$ will be zero. At any rate, we can introduce a nonsingular diagonal matrix $\mathcal{C}^\dagger \in \mathbb{R}^{p\times p}$, where the diagonal elements of $\mathcal{C}^\dagger$ correspond to $c_i^{-1}$ whenever $c_i \neq 0$ in the corresponding diagonal element in $\mathcal{C}$, and 1 elsewhere. We can also introduce the matrix $\mathcal{J} = \mathcal{C}^\dagger \mathcal{C} \in \mathbb{R}^{p\times m}$, which will have the same

structure as $\mathcal{C}$ with ones replacing the nonzero diagonal values in $\mathcal{C}$. With these notations, we can then write

$$\begin{aligned} P &= I_m - W\mathcal{C}^T\mathcal{C}^\dagger \left(\frac{1}{\alpha}\mathcal{C}^\dagger V_\gamma^T \Lambda_\gamma V_\gamma \mathcal{C}^\dagger + \mathcal{C}^\dagger \mathcal{C}\mathcal{C}^T\mathcal{C}^\dagger\right)^{-1} \mathcal{C}^\dagger \mathcal{C} W^T \\ &= I_m - W\mathcal{J}^T \left(\frac{1}{\alpha}\mathcal{C}^\dagger V_\gamma^T \Lambda_\gamma V_\gamma \mathcal{C}^\dagger + \mathcal{J}\mathcal{J}^T\right)^{-1} \mathcal{J} W^T, \end{aligned} \tag{6.8}$$

where the matrix $\mathcal{J}\mathcal{J}^T$ in the internal inverse operator is a $p \times p$ diagonal matrix with ones in places corresponding to the nonzero cosines in $\mathcal{C}$, and zeros elsewhere.

Consider now the most general case where $p \leq m$ and all cosines are nonzero, so that there will be no zero rows in the matrix $\mathcal{C}$,

$$\mathcal{C} = \mathrm{diag}(c_1, \ldots, c_p) \in \mathbb{R}^{p\times m}$$
$$= \left[\begin{array}{c:c} \begin{matrix} c_1 & & \\ & \ddots & \\ & & c_p \end{matrix} & 0 \end{array}\right],$$

with $c_i \neq 0,\ i = 1, \ldots, p$. This corresponds to the generic situation that one may encounter, assuming that the eigenvalues in $A$ are well clustered (after the first level of preconditioning) so that $p$ is small with respect to the number of constraints, and no orthogonality occurs between $\mathcal{I}m(U_\gamma)$ and $\mathcal{I}m(B)$. In this situation, $\mathcal{J}\mathcal{J}^T$ reduces to the identity matrix $I_p$, and $\mathcal{C}^\dagger$ can be written as $C_\gamma^{-1}$ where $C_\gamma \in \mathbb{R}^{p\times p}$ is the diagonal part of $\mathcal{C}$. Finally, let us denote by $W_\gamma = W\mathcal{J}^T \in \mathbb{R}^{m\times p}$ the subset of the $p$ first columns in $W$. The equation (6.8) then reduces to

$$P = I_m - W_\gamma \left(\frac{1}{\alpha} C_\gamma^{-1} V_\gamma^T \Lambda_\gamma V_\gamma C_\gamma^{-1} + I_p\right)^{-1} W_\gamma^T. \tag{6.9}$$

We can observe that the term $C_\gamma^{-1} V_\gamma^T \Lambda_\gamma V_\gamma C_\gamma^{-1}$ shows the relation between the cosines and the small eigenvalues of $A$. It is the key part of our analysis of interation through the Schur complement. Indeed, since $V_\gamma$ is orthogonal, we have

$$\begin{aligned} \frac{1}{\alpha}\|C_\gamma^{-1}(V_\gamma^T \Lambda_\gamma V_\gamma)C_\gamma^{-1}\|_2 &\leq \frac{1}{\alpha}\frac{\max\{\lambda_i\}_{i=1}^p}{(\min\{c_i\}_{i=1}^p)^2} \\ &\leq \frac{1}{\alpha}\frac{\gamma}{(\min\{c_i\}_{i=1}^p)^2}. \end{aligned}$$

From this inequality, we can see that the influence of the small eigenvalues in $\Lambda_\gamma$ (all those below $\gamma$) is inhibited in the inner inverse operator in (6.9) if

$$(\min\{c_i\}_{i=1}^p)^2 \gg \frac{\gamma}{\alpha}, \tag{6.10}$$

because in this case $\frac{1}{\alpha}\|C_\gamma^{-1}(V_\gamma^T\Lambda_\gamma V_\gamma)C_\gamma^{-1}\|_2 \ll 1$. Reasonable choices for $\gamma$ (for instance, $\gamma = \lambda_{\max}(A)/100$ and $\alpha = \lambda_{\max}(A)$) lead to values of $\gamma/\alpha \leq 10^{-2}$ so that we can expect an influence of the small eigenvalues only when there exist principal angles between $\mathcal{I}m(U_\gamma)$ and $\mathcal{I}m(B)$ whose cosines are less than $10^{-1}$. In this situation indeed $\frac{1}{\alpha}C_\gamma^{-1}(V_\gamma^T\Lambda_\gamma V_\gamma)C_\gamma^{-1}$ influences the identity matrix $I_p$ in (6.9).

## 6.2 Illustrations

In the previous section, we highlighted the influence of the cosines values of the principal angles between $\mathcal{I}m(B)$ and $\mathcal{I}m(U_\gamma)$ on the approximation of the inverse of the Schur complement. In the situation where the small eigenvalues of $A$ can have an influence on the Schur complement, the preconditionner

$$\mathcal{P}_1 = \begin{bmatrix} A_\gamma & 0 \\ 0 & S_\gamma \end{bmatrix}, \tag{6.11}$$

introduced in Chapter 4, has a sizeable impact on the convergence of MINRES and we present some illustrations showing this effect in the next sections.

### 6.2.1 On a toy example

In this section, we use a toy example to show that these cosines values can impact and spoil the convergence of MINRES. To build the matrix $\mathcal{A}_{KKT}$, we consider a diagonal matrix $A$ of order $n = 500$ with diagonal entries in $]0, 1]$ and such that

$$A = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix},$$

with $A_1 = \Lambda_\gamma \in \mathbb{R}^{p\times p}$ where $p = 5$ and

$$\begin{aligned} \Lambda_\gamma &= \operatorname{diag}(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5) \\ &= \operatorname{diag}(10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}, 10^{-1}), \end{aligned}$$

and $A_2 \in \mathbb{R}^{(n-p)\times(n-p)}$, where $n - p = 495$, is a diagonal matrix with uniform values from the interval $[a, b] = [0.11, 1]$ and randomly generated by the `Matlab` code

```
diag(a + (b-a).*rand(n-p,1)).
```

We impose that the columns of the matrix $U_\gamma$ are the first $p$ vectors $\{e_1, e_2, \ldots, e_p\}$ of the canonical basis of $\mathbb{R}^n$. The matrix $B \in \mathbb{R}^{n\times m}$, where $m = 200$, is set to

$$B = \begin{bmatrix} C_\gamma & 0 \\ B_1 S_\gamma & B_2 \end{bmatrix},$$

where $C_\gamma = \mathrm{diag}\{\ \cos\theta_i\ \}_{i=1}^p$, $S_\gamma = \mathrm{diag}\{\ \sin\theta_i\ \}_{i=1}^p$, $B_1 \in \mathbb{R}^{(n-p)\times p}$, $B_2 \in \mathbb{R}^{(n-p)\times(m-p)}$, that are such that $Q = [B_1 \quad B_2] \in \mathbb{R}^{(n-p)\times m}$ satisfies $Q^TQ = I_m$ ensuring that $B$ has orthonormal columns. Indeed, we have

$$\begin{aligned} B^TB &= \begin{bmatrix} C_\gamma & S_\gamma B_1^T \\ 0 & B_2^T \end{bmatrix} \begin{bmatrix} C_\gamma & 0 \\ B_1 S_\gamma & B_2 \end{bmatrix} \\ &= \begin{bmatrix} C_\gamma^2 + S_\gamma B_1^T B_1 S_\gamma & 0 \\ 0 & B_2^T B_2 \end{bmatrix} \\ &= \begin{bmatrix} I_p & 0 \\ 0 & I_{m-p} \end{bmatrix}. \end{aligned}$$

We then obtain

$$\begin{aligned} K_\gamma &= (B^TB)^{-1/2} B^T U_\gamma \\ &= B^T U_\gamma \\ &= \begin{bmatrix} C_\gamma & S_\gamma B_1^T \\ 0 & B_2^T \end{bmatrix} \begin{bmatrix} I_p \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} C_\gamma \\ 0 \end{bmatrix}, \end{aligned}$$

which corresponds to the CS decomposition (6.5) of $K_\gamma^T$ with the matrices $V_\gamma$ and $W$ equal to the identity matrix. The principal vectors in $\mathcal{I}m(U_\gamma)$ are equal to $U_\gamma V_\gamma = U_\gamma$ and the rank-$p$ update

$$\frac{1}{\alpha} C_\gamma^{-1} (V_\gamma^T \Lambda_\gamma V_\gamma) C_\gamma^{-1} + I_p$$

of $S_\gamma^{-1}$ in (6.9) is reduced to

$$\frac{1}{\alpha} C_\gamma^{-1} \Lambda_\gamma C_\gamma^{-1} + I_p.$$

It implies a one-to-one match between eigenvectors versus principal vectors and eigenvalues below $\gamma$ versus cosines of the principal angles between $\mathcal{I}m(U_\gamma)$ and $\mathcal{I}m(B)$. We consider three different configurations $(a), (b)$ and $(c)$ for

$$C_\gamma = \mathrm{diag}\{\cos\theta_i\}_{i=1}^5,$$

with values of the cosines of the principal angles between $\mathcal{I}m(U_\gamma)$ and $\mathcal{I}m(B)$ given in Table 6.1.

| | $(a)$ | $(b)$ | $(c)$ |
|---|---|---|---|
| $\cos\theta_1$ | 0.3 | 0.3 | $10^{-6}$ |
| $\cos\theta_2$ | 0.3 | $10^{-3}$ | $10^{-5}$ |
| $\cos\theta_3$ | 0.3 | 0.3 | $10^{-4}$ |
| $\cos\theta_4$ | 0.3 | 0.3 | $10^{-3}$ |
| $\cos\theta_5$ | 0.3 | 0.3 | 0.3 |

Table 6.1 – Values of the cosines of the principal angles between $\mathcal{I}m(U_\gamma)$ and $\mathcal{I}m(B)$ for three configurations.

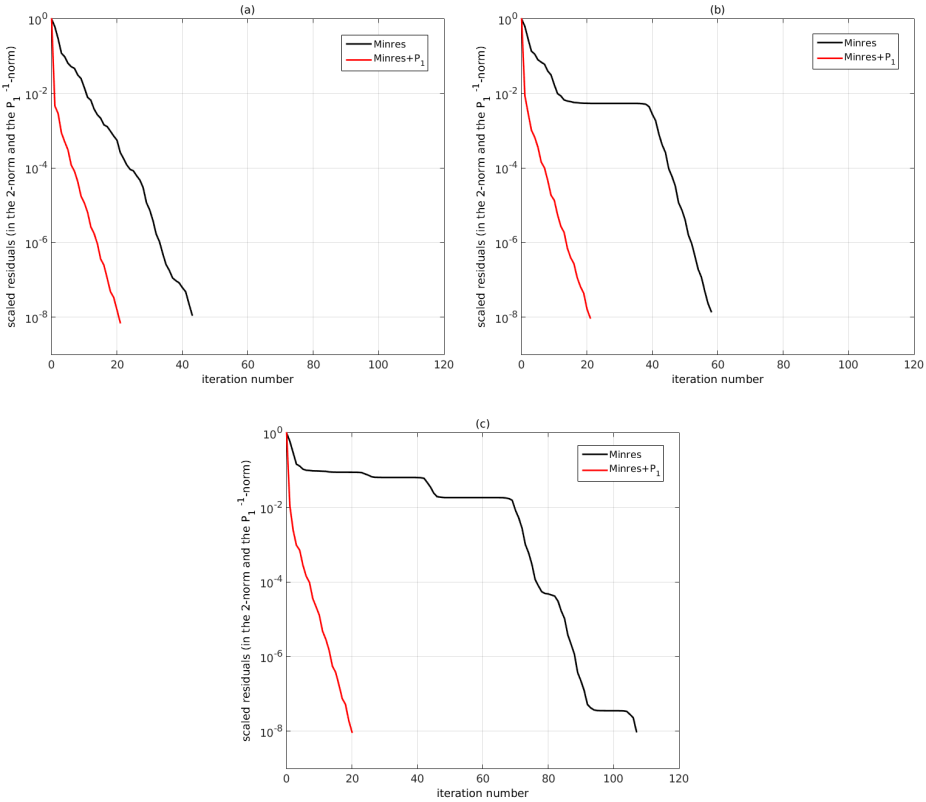



Figure 6.1 – Convergence profiles (2-norm and $\mathcal{P}_1^{-1}$-norm of relative residuals) for different values of $C_\gamma$.

Figure 6.1 illustrates and compares the impact of the values of the cosines of the principal angles between $\mathcal{I}m(U_\gamma)$ and $\mathcal{I}m(B)$ on the behaviour of MINRES applied to our toy KKT matrix for the three cases. Furthermore, we show the impact of the preconditioner $\mathcal{P}_1$. For comparison purposes, the iterations are stopped when the scaled residual in 2-norm or $\mathcal{P}_1^{-1}$ is less than $10^{-8}$. For particular values of the cosines, the phenomenon of plateau occurs. Indeed, we come back to the relation $\frac{1}{\alpha}C_\gamma^{-1}\Lambda_\gamma C_\gamma^{-1} + I_p$, which implies that if the square of the inverse of the cosines of some principal angles are equal to the corresponding eigenvalues, the corresponding bad conditioning of $A$ is showed up in the Schur complement inverse. For instance, if we change the value of the second cosine $\cos\theta_2$ from 0.3 to $10^{-3}$ (corresponding to the square root of the corresponding eigenvalue) between situations $(a)$ and $(b)$, the speed of convergence of preconditioned MINRES by $\mathcal{P}_1^{-1}$ is disrupted. One such phenomenon of plateau in the convergence curve occurs in Figure 6.1, case $(b)$. After this phenomenon of plateau, the convergence behaviour is similar to the classical one for MINRES. Case $(c)$ corresponds to a generalized case where four out of the five values of the cosines reveal the corresponding bad conditioning of $A$ in the Schur complement inverse which leads to four phenomena of plateau in Figure 6.1, case $(c)$. In all situations, we have that the convergence of preconditioned MINRES by $\mathcal{P}_1^{-1}$, which deals with the bad conditioning of $A$, is linear and we observe that the number of iterations in all situations is constant.

### 6.2.2 Varying the constraint matrix

The analysis in this section tells us that if all the cosines of the principal angles between $\mathcal{I}m(U_\gamma)$ and $\mathcal{I}m(B)$ are large enough, the convergence of MINRES preconditioned with the classical preconditioner (4.18),

$$\mathcal{P}_{IBB} = \begin{bmatrix} I_n & 0 \\ 0 & B^TB \end{bmatrix},$$

should be fast, independently of any consideration with respect to the ill-conditioning in $A$. To illustrate this case, we consider a KKT matrix with the same matrix $A$ as in Section 3.1, with a choice of $\gamma = \lambda_{\max}(A)/100 \approx 3.8\,10^{-2}$ and $\alpha = 1.16$. The dimension of the invariant subspace $\mathcal{I}m(U_\gamma)$ is $p = 42$. As we have seen before, the influence of the small eigenvalues in $\Lambda_\gamma$ is inhibited in the inner inverse operator in (6.9) if

$$(\min\{c_i\}_{i=1}^p)^2 \gg \frac{\gamma}{\alpha},$$

or, equivalently,

$$\min\{c_i\}_{i=1}^p \gg \sqrt{\frac{\gamma}{\alpha}}. \tag{6.12}$$

We thus change the constraint matrix $B$ to $\widetilde{B}$ so as to get the cosines of the principal angles between $\mathcal{I}m(U_\gamma)$ and $\mathcal{I}m(\widetilde{B})$ to be larger than $2\sqrt{\gamma/\alpha} \simeq 0.362$,

which affects the $\ell = 22$ smallest cosines in this example. By (6.4), we have that

$$\begin{aligned} B &= Q(B^TB)^{1/2} \\ &= QW\ W^T(B^TB)^{1/2} \\ &= [QW_\ell,\, QW_{\not\ell}]\ W^T(B^TB)^{1/2} \end{aligned} \tag{6.13}$$

where $Q \in \mathbb{R}^{n\times m}$ satisfies $Q^TQ = I_m$, $W \in \mathbb{R}^{m\times m}$ denotes an orthogonal matrix introduced in the CS decomposition of $K^T$ given by (6.5)-(6.7), $W_\ell \in \mathbb{R}^{m\times \ell}$ denotes the subset of columns in $W$ associated to the selected smallest cosines, and $W_{\not\ell}$ denotes the submatrix in $W$ made with the remaining $(m-\ell)$ columns not included in $W_\ell$. We change $B$ into

$$\widetilde{B} = [QW_\ell\Omega_\ell + U_\gamma V_\ell \Phi_\ell,\, QW_{\not\ell}]\, W^T(B^TB)^{1/2}, \tag{6.14}$$

where the diagonal square matrices $\Omega_\ell = \mathrm{diag}(\omega_i)_{1\le i\le \ell}$ and $\Phi_\ell = \mathrm{diag}(\varphi_i)_{1\le i\le \ell}$, and where $V_\ell \in \mathbb{R}^{p\times \ell}$ denotes the submatrix made with the columns from $V_\gamma$ associated to the selected smallest cosines. Using the notations

$$Y = U_\gamma V_\gamma \in \mathcal{I}m(U_\gamma) \quad \text{and} \quad Z = QW \in \mathcal{I}m(B),$$

we have that the $p$ first columns of the matrix $Y$ and $Z$ are the $p$ principal vectors (see (6.5) and (6.6) with $p \le m$ and $p \le n-p$ implying that $\min\{r, q\} = p$). Observe that, the principal vectors $QW_\ell$ in (6.13) are modified in (6.14) as a linear combination between the principal vectors $QW_\ell$ and $U_\gamma V_\ell$ associated to the selected cosines. The choice of the diagonal entries $\omega_i$ and $\varphi_i$, $1 \le i \le \ell$, is made so that $\widetilde{B}(B^TB)^{-1/2}$ still corresponds to a set of $m$ orthonormal vectors, i.e.,

$$\|\tilde{z}_i\|_2^2 = 1,$$

or equivalently,

$$\omega_i^2 + \varphi_i^2 + 2\omega_i\varphi_i c_i = 1, \tag{6.15}$$

where $\tilde{z}_i = \omega_i z_i + \varphi_i y_i$ with $z_i$ and $y_i$ are the $i$th columns of $Y$ and $Z$, respectively and $c_i = z_i^T y_i$. Furthermore, we impose that the cosines of the principal angles between $\mathcal{I}m(U_\gamma)$ and $\mathcal{I}m(\widetilde{B})$ determined by

$$\tilde{c}_i = \tilde{z}_i^T y_i,$$

or, equivalently,

$$\tilde{c}_i = \omega_i c_i + \varphi_i, \tag{6.16}$$

are now driven to a specific predetermined value $\tilde{c}_i = 0.362$, for $i = 1, \ldots, \ell$. By (6.16), we have

$$\varphi_i = \tilde{c_i} - \omega_i c_i, \tag{6.17}$$

and substituting in (6.15), we obtain

$$\omega_i^2 + (\tilde{c_i} - \omega_i c_i)^2 + 2\omega_i(\tilde{c_i} - \omega_i c_i)c_i = 1, \tag{6.18}$$

and thus

$$\omega_i^2 + \tilde{c_i}^2 - \omega_i^2 c_i^2 = 1, \tag{6.19}$$

leading to

$$\omega_i = \sqrt{\frac{1-\tilde{c}_i^2}{1-c_i^2}} \quad \text{and} \quad \varphi_i = \tilde{c}_i - \omega_i c_i\,, \quad \text{for} \quad i = 1, \ldots, \ell.$$

Note that by (6.14), in this way,

$$\begin{aligned}(\widetilde{B}^T\widetilde{B})^{-1/2} &= \left(\left((B^TB)^{1/2}W\right) X \left(W^T(B^TB)^{1/2}\right)\right)^{-1/2} \\ &= (B^TB)^{-1/2}\end{aligned}$$

where

$$\begin{aligned}X &= \begin{bmatrix} \Omega_\ell W_\ell^T Q^T + \Phi_\ell V_\ell^T U_\gamma^T \\ W_{\not\ell}^T Q^T \end{bmatrix} \begin{bmatrix} QW_\ell\Omega_\ell + U_\gamma V_\ell \Phi_\ell & QW_{\not\ell} \end{bmatrix} \\ &= \begin{bmatrix} \left(\Omega_\ell W_\ell^T Q^T + \Phi_\ell V_\ell^T U_\gamma^T\right)\left(QW_\ell\Omega_\ell + U_\gamma V_\ell\Phi_\ell\right) & \Omega_\ell W_\ell^T W_{\not\ell} + \Phi_\ell V_\ell^T U_\gamma^T Q W_{\not\ell} \\ W_{\not\ell}^T W_\ell \Omega_\ell + W_{\not\ell}^T Q^T U_\gamma V_\ell \Phi_\ell & W_{\not\ell}^T W_{\not\ell} \end{bmatrix} \\ &= I_m,\end{aligned}$$

and that $\widetilde{B}$ incorporates the same ill-conditioning as the one of matrix $B$.

Table 6.2 gives the $\ell = 22$ smallest cosines $c_i$ and the corresponding values for $\omega_i$ and $\varphi_i$ to drive all these cosines to the fixed value $\tilde{c}_i = 0.362$. We can observe that the scalar values $\varphi_i$ are small with respect to that of $\omega_i$, and that the perturbation added to $B$ is relatively small since $QW_\ell$ in (6.13) has been replaced by $QW_\ell\Omega_\ell + U_\gamma V_\ell\Phi_\ell$ in (6.14). Nevertheless, this is enough to set up MINRES in good conditions for linear convergence with a preconditioner taking care of the bad conditioning in the constraints only. Figure 6.2 shows the convergence profiles of MINRES preconditioned with $\mathcal{P}_{IBB}$ as well as with $\mathcal{P}_1$ in (6.11) (no convergence without preconditioning). We can observe that the use of the spectral information incorporated in $\mathcal{P}_1$ does not drastically improve the situation in this case. The interference (or recombination) between $A$ and $B$ has almost no impact in this situation with large enough cosines. This illustrates the analysis and comments made above, which also extends

| $c_i$ | $\omega_i$ | $\varphi_i$ | $\tilde{c}_i$ |
|---|---|---|---|
| $4.98\,10^{-3}$ | $9.32\,10^{-1}$ | $3.58\,10^{-1}$ | $3.62\,10^{-1}$ |
| $6.88\,10^{-3}$ | $9.32\,10^{-1}$ | $3.56\,10^{-1}$ | $3.62\,10^{-1}$ |
| $8.21\,10^{-3}$ | $9.32\,10^{-1}$ | $3.55\,10^{-1}$ | $3.62\,10^{-1}$ |
| $1.23\,10^{-2}$ | $9.32\,10^{-1}$ | $3.51\,10^{-1}$ | $3.62\,10^{-1}$ |
| $1.72\,10^{-2}$ | $9.32\,10^{-1}$ | $3.46\,10^{-1}$ | $3.62\,10^{-1}$ |
| $2.39\,10^{-2}$ | $9.32\,10^{-1}$ | $3.40\,10^{-1}$ | $3.62\,10^{-1}$ |
| $3.03\,10^{-2}$ | $9.32\,10^{-1}$ | $3.34\,10^{-1}$ | $3.62\,10^{-1}$ |
| $4.00\,10^{-2}$ | $9.33\,10^{-1}$ | $3.25\,10^{-1}$ | $3.62\,10^{-1}$ |
| $4.32\,10^{-2}$ | $9.33\,10^{-1}$ | $3.22\,10^{-1}$ | $3.62\,10^{-1}$ |
| $4.62\,10^{-2}$ | $9.33\,10^{-1}$ | $3.19\,10^{-1}$ | $3.62\,10^{-1}$ |
| $5.51\,10^{-2}$ | $9.33\,10^{-1}$ | $3.11\,10^{-1}$ | $3.62\,10^{-1}$ |
| $6.05\,10^{-2}$ | $9.34\,10^{-1}$ | $3.06\,10^{-1}$ | $3.62\,10^{-1}$ |
| $7.20\,10^{-2}$ | $9.34\,10^{-1}$ | $2.95\,10^{-1}$ | $3.62\,10^{-1}$ |
| $7.83\,10^{-2}$ | $9.35\,10^{-1}$ | $2.89\,10^{-1}$ | $3.62\,10^{-1}$ |
| $8.86\,10^{-2}$ | $9.36\,10^{-1}$ | $2.80\,10^{-1}$ | $3.62\,10^{-1}$ |
| $1.46\,10^{-1}$ | $9.42\,10^{-1}$ | $2.25\,10^{-1}$ | $3.62\,10^{-1}$ |
| $1.49\,10^{-1}$ | $9.43\,10^{-1}$ | $2.22\,10^{-1}$ | $3.62\,10^{-1}$ |
| $1.69\,10^{-1}$ | $9.46\,10^{-1}$ | $2.03\,10^{-1}$ | $3.62\,10^{-1}$ |
| $2.02\,10^{-1}$ | $9.52\,10^{-1}$ | $1.71\,10^{-1}$ | $3.62\,10^{-1}$ |
| $2.83\,10^{-1}$ | $9.72\,10^{-1}$ | $8.77\,10^{-2}$ | $3.62\,10^{-1}$ |
| $3.22\,10^{-1}$ | $9.85\,10^{-1}$ | $4.52\,10^{-2}$ | $3.62\,10^{-1}$ |
| $3.33\,10^{-1}$ | $9.88\,10^{-1}$ | $3.32\,10^{-2}$ | $3.62\,10^{-1}$ |

Table 6.2 – Values of the $\ell = 22$ cosines of the principal angles between $\mathcal{I}m(U_\gamma)$ and $\mathcal{I}m(B)$, and corresponding coefficients $\omega_i$ and $\varphi_i$ for the linear combination of the associated principal vectors to achieve the target cosine value $\tilde{c}_i$.

the results in Rusten and Winther (1992) when $\sigma_i = 1,\ i = 1\ldots,m$ and $\lambda_1 \gg 0$ that clearly imply that the convergence of Minres must be fast when the $(1,1)$ block is well-conditioned and the constraint matrix is for instance orthogonalized (as done with $\mathcal{P}_{IBB}$).

Considering preconditioners $\mathcal{P}_1$ and $\mathcal{P}_{IBB}$, Table 6.3 provides the number of iterations of preconditioned Minres on $\mathcal{A}_{KKT}x = b$ with, in $\mathcal{A}_{KKT}$, $B$ or $\tilde{B}$ as described above. We can see that the preconditioner $\mathcal{P}_1$ significantly decreases the number of iterations with respect to $\mathcal{P}_{IBB}$ when we consider $B$ with cosines of principal angles between $\mathcal{I}m(U_\gamma)$ and $\mathcal{I}m(B)$ less than $2\sqrt{\gamma/\alpha}$ .

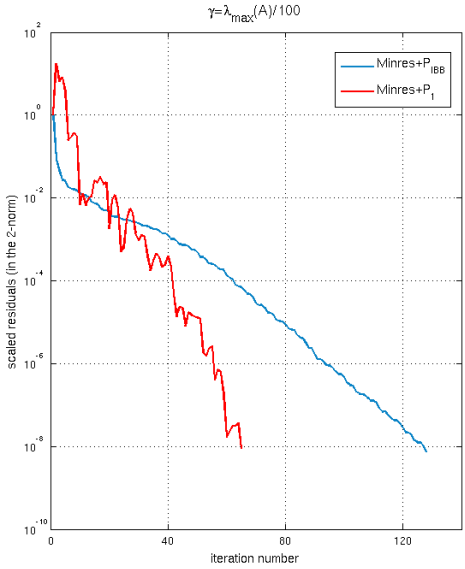



Figure 6.2 – Convergence profiles of preconditioned MINRES (with $\mathcal{P}_{IBB}$ and $\mathcal{P}_1$) in the case of large enough principal angles cosines.

| $\mathcal{P}$ | # iter. MINRES for system with $B$ | # iter. MINRES for system with $\tilde{B}$ |
|---|---|---|
| $\mathcal{P}_1$ | 62 | 65 |
| $\mathcal{P}_{IBB}$ | 483 | 128 |

Table 6.3 – Number of iterations of preconditioned MINRES on the system of matrix $\mathcal{A}_{KKT}$ with $B$ or $\tilde{B}$.

## 6.3 Refined eigenvalue bounds for KKT matrices

Consider the KKT matrix,

$$\mathcal{A}_{KKT} = \begin{bmatrix} A & Q \\ Q^T & 0 \end{bmatrix}, \tag{6.20}$$

with $Q^TQ = I_m$ (corresponding to a constraint matrix $B$ whose columns are orthonormal) and assume that a first level of preconditioning has been applied so that the largest eigenvalue of $A$ is equal to one. Using the fundamental result from Rusten and Winther (1992), see Theorem 2.5 in Chapter 2, the

eigenvalues of $\mathcal{A}_{KKT}$ are bounded within the intervals

$$\left[\frac{\lambda_{\min}(A) - \sqrt{\lambda_{\min}^2(A)+4}}{2}, \frac{1-\sqrt{5}}{2}\right] \bigcup \left[\lambda_{\min}(A), \frac{1+\sqrt{5}}{2}\right], \quad (6.21)$$

since $\lambda_{\max}(A) = 1$ and the singular values of $Q$ are equal to 1 as well. The left interval in (6.21), associated to the negative eigenvalues in $\mathcal{A}_{KKT}$, is basically well bounded and isolated away from zero, as opposed to the right interval (the one associated to the positive eigenvalues in $\mathcal{A}_{KKT}$) which is well bounded towards infinity but not isolated away from zero. The smallest eigenvalue $\lambda_{\min}(A)$ could possibly tend to zero.

In this section, we aim at refining the lower bound $\lambda_{\min}(A)$ of the right interval in (6.21) through a theoretical analysis in terms of cosines of principal angles between $\mathcal{I}m(U_\gamma)$ and $\mathcal{I}m(Q)$. Doing so, we expect to identify situations where this lower bound is guaranteed bounded away from zero. In Section 6.3.1, we first deduce some general spectral relations and we focus on the eigenvalues of the matrix in (6.20) smaller than $\gamma/2$ in Section 6.3.2. Finally, based on these spectral relations, we successively define two constrained optimization problems that will lead to the desired result of refining the positive lower bound of (6.21). The norm considered in this section is the 2-norm $\|\cdot\|_2$ and we use in the following, the short notation $\|\cdot\|$.

### 6.3.1 General spectral relations

We rewrite the matrix (6.20) into two successively similar matrices. We first consider the eigendecomposition (3.3) of the matrix $A$

$$A = U\Lambda U^T,$$

where the diagonal matrix $\Lambda \in \mathbb{R}^{n\times n}$ contains the eigenvalues $\{\lambda_i\}_{i=1}^n$ of $A$ and the orthonormal matrix $U \in \mathbb{R}^{n\times n}$ contains the associated orthonormal eigenvectors. We first observe that

$$\begin{aligned}\begin{bmatrix} A & Q \\ Q^T & 0 \end{bmatrix} &= \begin{bmatrix} U\Lambda U^T & Q \\ Q^T & 0 \end{bmatrix} \\ &= \begin{bmatrix} U & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \Lambda & U^TQ \\ Q^TU & 0 \end{bmatrix} \begin{bmatrix} U^T & 0 \\ 0 & I \end{bmatrix}. \end{aligned} \quad (6.22)$$

We next split the spectrum of $A$ in two parts (similarly to Section 3.1), with $\Lambda_\gamma \in \mathbb{R}^{p\times p}$, the diagonal matrix containing the $p$ eigenvalues $0 < \lambda_1 \leq \ldots \leq \lambda_p$ less or equal than a given positive number $\gamma \in [\lambda_{\min}(A), 1]$, and with $\tilde{\Lambda}_\gamma \in \mathbb{R}^{(n-p)\times(n-p)}$ the diagonal matrix containing all the other $(n-p)$ eigenvalues $\lambda_{p+1} \leq \ldots \leq \lambda_n$. The matrix $U = [U_\gamma,\ \tilde{U}_\gamma] \in \mathbb{R}^{n\times n}$ is orthogonal, where the columns of the rectangular matrices $U_\gamma \in \mathbb{R}^{n\times p}$ and $\tilde{U}_\gamma \in \mathbb{R}^{n\times(n-p)}$ are

the orthonormal sets of eigenvectors corresponding to $\Lambda_\gamma$ and $\tilde{\Lambda}_\gamma$, respectively. $U = [U_\gamma,\ \tilde{U}_\gamma] \in \mathbb{R}^{n\times n}$. Based on the CS decomposition of $K^T$ introduced in Section 6.1, (6.5), (6.6) and the relation (6.7) implies without loss of generality that, if $p < m$ and $m < n - p$, then $r = p$ and $q = m$, so that

$$\mathcal{C} = \begin{bmatrix} C & 0 \end{bmatrix} \in \mathbb{R}^{p\times m} \quad \text{and} \quad \mathcal{S} = \begin{bmatrix} S & 0 \\ 0 & I_{m-p} \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{(n-p)\times m} \tag{6.23}$$

with $C \in \mathbb{R}^{p\times p}$ and $S \in \mathbb{R}^{p\times p}$. Extracting $K_\gamma$ and $\tilde{K}_\gamma$ from (6.5) and (6.6) yields

$$K_\gamma = W\mathcal{C}^T V_\gamma^T \qquad \text{and} \qquad \tilde{K}_\gamma = W\mathcal{S}^T \tilde{V}_\gamma^T. \tag{6.24}$$

Coming back to the matrix in (6.22), using the expressions (6.24) and remembering that $K = [K_\gamma,\ \tilde{K}_\gamma]$ by (6.3), we rewrite this matrix in terms of cosines and sines, as

$$\begin{bmatrix} \Lambda_\gamma & 0 & V_\gamma \mathcal{C} W^T \\ 0 & \tilde{\Lambda}_\gamma & \tilde{V}_\gamma \mathcal{S} W^T \\ W\mathcal{C}^T V_\gamma^T & W\mathcal{S}^T \tilde{V}_\gamma^T & 0 \end{bmatrix}.$$

We next obtain

$$\begin{bmatrix} V_\gamma^T & 0 & 0 \\ 0 & \tilde{V}_\gamma^T & 0 \\ 0 & 0 & W^T \end{bmatrix} \begin{bmatrix} \Lambda_\gamma & 0 & V_\gamma \mathcal{C} W^T \\ 0 & \tilde{\Lambda}_\gamma & \tilde{V}_\gamma \mathcal{S} W^T \\ W\mathcal{C}^T V_\gamma^T & W\mathcal{S}^T \tilde{V}_\gamma^T & 0 \end{bmatrix} \begin{bmatrix} V_\gamma & 0 & 0 \\ 0 & \tilde{V}_\gamma & 0 \\ 0 & 0 & W \end{bmatrix} = \begin{bmatrix} M_\gamma & 0 & \mathcal{C} \\ 0 & \tilde{M}_\gamma & \mathcal{S} \\ \mathcal{C}^T & \mathcal{S}^T & 0 \end{bmatrix} \tag{6.25}$$

where

$$M_\gamma = V_\gamma^T \Lambda_\gamma V_\gamma \quad \text{and} \quad \tilde{M}_\gamma = \tilde{V}_\gamma^T \tilde{\Lambda}_\gamma \tilde{V}_\gamma, \tag{6.26}$$

by the orthonormality of the columns of $V_\gamma, \tilde{V}_\gamma$ and $W$. Replacing (6.23) in (6.25), then yields the matrix

$$\left[\begin{array}{c|ccc:c:c} M_\gamma & \multicolumn{3}{c:}{0} & C & 0 \\ \hline & & & & S & 0 \\ 0 & \multicolumn{3}{c:}{\tilde{M}_\gamma} & 0 & I \\ & & & & 0 & 0 \\ \hline C & S & 0 & 0 & 0 & 0 \\ \hdashline 0 & 0 & I & 0 & 0 & 0 \end{array}\right]. \tag{6.27}$$

Let $\nu \in \mathbb{R}$ denote an eigenvalue of this matrix, with $\begin{bmatrix} x_1 & x_2 & y_1 & y_2 \end{bmatrix}^T$ where $x_1 \in \mathbb{R}^p$, $x_2 \in \mathbb{R}^{n-p}$, $y_1 \in \mathbb{R}^p$ and $y_2 \in \mathbb{R}^{m-p}$, the associated eigenvector, from

$$\left[\begin{array}{c|cc|c:c} M_\gamma & 0 & & C & 0 \\ \hline & & & S & 0 \\ 0 & & \tilde{M}_\gamma & 0 & I \\ & & & 0 & 0 \\ \hline C & S & 0 \; 0 & 0 & 0 \\ \hdashline 0 & 0 & I \; 0 & 0 & 0 \end{array}\right] \begin{bmatrix} x_1 \\ x_2 \\ y_1 \\ y_2 \end{bmatrix} = \nu \begin{bmatrix} x_1 \\ x_2 \\ y_1 \\ y_2 \end{bmatrix}, \tag{6.28}$$

that

$$\begin{aligned} M_\gamma x_1 + C y_1 &= \nu x_1, && (6.29) \\ \tilde{M}_\gamma x_2 + \begin{bmatrix} S y_1 \\ y_2 \\ 0 \end{bmatrix} &= \nu x_2, && (6.30) \\ C x_1 + \begin{bmatrix} S & 0 & 0 \end{bmatrix} x_2 &= \nu y_1, && (6.31) \end{aligned}$$

and

$$\begin{bmatrix} 0 & I & 0 \end{bmatrix} x_2 = \nu y_2. \tag{6.32}$$

Let $\rho_1$ and $\rho_2 \in \mathbb{R}$ be positive quantities satisfying

$$\begin{aligned} \rho_1 \in [\lambda_1, \lambda_p] \quad \text{and} \quad \rho_2 \in [\lambda_{p+1}, 1], && (6.33) \\ x_1^T M_\gamma x_1 = \rho_1 \|x_1\|^2, && (6.34) \end{aligned}$$

and

$$x_2^T \tilde{M}_\gamma x_2 = \rho_2 \|x_2\|^2. \tag{6.35}$$

Note that if both $x_1 \neq 0$ and $x_2 \neq 0$, then $\rho_1$ and $\rho_2$ are Rayleigh quotients and satisfy $\rho_1 \in [\lambda_1, \lambda_p]$ and $\rho_2 \in [\lambda_{p+1}, 1]$ by (6.26), respectively, implying (6.33). Otherwise, if $x_1 = 0$ or $x_2 = 0$ or both, it is always possible to find positive quantities $\rho_1$ and $\rho_2$ satisfying (6.33), (6.34) and (6.35).

The following lemma gives some spectral relations, which will be useful in the next sections.

**Lemma 6.1** Let $\nu \in \mathbb{R}$ be an eigenvalue of the matrix (6.27) associated to the eigenvector $x = \begin{bmatrix} x_1 & x_2 & y_1 & y_2 \end{bmatrix}^T$ with $x_1 \in \mathbb{R}^p$, $x_2 \in \mathbb{R}^{n-p}$, $y_1 \in \mathbb{R}^p$ and $y_2 \in \mathbb{R}^{m-p}$. Then $x$ satisfies the following relations

$$\nu^2\|x_1\|^2 = \nu\rho_1\|x_1\|^2 + x_1^T C^2 x_1 + x_1^T \begin{bmatrix} CS & 0 & 0 \end{bmatrix} x_2, \tag{6.36}$$

$$\nu^2\|x_2\|^2 = \nu\rho_2\|x_2\|^2 + x_1^T \begin{bmatrix} CS & 0 & 0 \end{bmatrix} x_2 + x_2^T \begin{bmatrix} S^2 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix} x_2, \tag{6.37}$$

$$\nu^2\|y_1\|^2 = x_1^T C^2 x_1 + 2x_1^T \begin{bmatrix} CS & 0 & 0 \end{bmatrix} x_2 + x_2^T \begin{bmatrix} S^2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} x_2, \tag{6.38}$$

$$\nu^2\|y_2\|^2 = x_2^T \begin{bmatrix} 0 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix} x_2, \tag{6.39}$$

where $\rho_1$ and $\rho_2$ satisfy (6.33), (6.34) and (6.35).

**Proof.** Multiplying (6.29) by $\nu x_1^T$ and using (6.34), we have

$$\begin{aligned} \nu^2\|x_1\|^2 &= \nu x_1^T M_\gamma x_1 + \nu x_1^T C y_1 \\ &= \nu\rho_1\|x_1\|^2 + \nu x_1^T C y_1. \end{aligned} \tag{6.40}$$

Multiplying now (6.31) by $x_1^T C$, we obtain

$$\nu x_1^T C y_1 = x_1^T C^2 x_1 + x_1^T \begin{bmatrix} CS & 0 & 0 \end{bmatrix} x_2,$$

which together with (6.40) implies (6.36). Similiarly, multiplying (6.30) by $\nu x_2^T$ and using (6.35), we obtain

$$\begin{aligned} \nu^2\|x_2\|^2 &= \nu x_2^T \tilde{M}_\gamma x_2 + \nu x_2^T \begin{bmatrix} S \\ 0 \\ 0 \end{bmatrix} y_1 + \nu x_2^T \begin{bmatrix} 0 \\ I \\ 0 \end{bmatrix} y_2 \\ &= \nu\rho_2\|x_2\|^2 + \nu x_2^T \begin{bmatrix} S \\ 0 \\ 0 \end{bmatrix} y_1 + \nu x_2^T \begin{bmatrix} 0 \\ I \\ 0 \end{bmatrix} y_2. \end{aligned}$$

Using (6.31) and (6.32) for $\nu y_1$ and $\nu y_2$, we derive

$$\nu^2\|x_2\|^2 = \nu\rho_2\|x_2\|^2 + x_2^T \begin{bmatrix} S \\ 0 \\ 0 \end{bmatrix} \left(Cx_1 + \begin{bmatrix} S & 0 & 0 \end{bmatrix} x_2\right) + x_2^T \begin{bmatrix} 0 \\ I \\ 0 \end{bmatrix} \begin{bmatrix} 0 & I & 0 \end{bmatrix} x_2,$$

implying (6.37). Finally, (6.38) and (6.39) immediately follow from (6.31) and (6.32), respectively. □

### 6.3.2 Spectral relations for small positive eigenvalues

As we have seen in (6.21), the lower bound of the interval associated to the positive eigenvalues in $\mathcal{A}_{KKT}$ given in (6.20) is not necessarily isolated away from zero. In this section, we thus focus our analysis on the eigenvalues of $\mathcal{A}_{KKT}$ smaller than $\gamma$ and we deduce, from Theorem 6.1, spectral relations associated to these eigenvalues. Let us assume that the eigenvalue problem (6.28) has an eigenvalue denoted by $\bar{\nu} > 0$ such that $\bar{\nu} < \gamma/2$ with the corresponding eigenvector $\bar{x} := \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \bar{y}_1 & \bar{y}_2 \end{bmatrix}^T$. Lemma 6.2 shows that the vectors $\bar{x}_1 \in \mathbb{R}^n$ and $\bar{x}_2 \in \mathbb{R}^{n-p}$ are nonzero due to the existence of such an eigenvalue $\bar{\nu}$.

**Lemma 6.2** Assume that the matrix (6.27) has an eigenvalue $\bar{\nu}$ satisfying $0 < \bar{\nu} < \gamma/2$, with the associated eigenvector $\bar{x} = \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \bar{y}_1 & \bar{y}_2 \end{bmatrix}^T$ where $\bar{x}_1 \in \mathbb{R}^p$, $\bar{x}_2 \in \mathbb{R}^{n-p}$, $\bar{y}_1 \in \mathbb{R}^p$ and $\bar{y}_2 \in \mathbb{R}^{m-p}$. Let also $C$ and $S \in \mathbb{R}^{p\times p}$ given by (6.23) satisfy $c_{\min} := \min_{i=1:p}\{c_i\} > 0$. Then, the vectors $\bar{x}_1$ and $\bar{x}_2$ are nonzero.

**Proof.** Let first show that $\bar{x}_1$ is nonzero. Summing (6.36) and (6.37), we have that

$$\begin{aligned} \bar{\nu}^2(\|\bar{x}_1\|^2 + \|\bar{x}_2\|^2) &= \bar{\nu}(\bar{\rho}_1\|\bar{x}_1\|^2 + \bar{\rho}_2\|\bar{x}_2\|^2) + \bar{x}_1^T C^2 \bar{x}_1 \\ &+ 2\bar{x}_1^T \begin{bmatrix} CS & 0 & 0 \end{bmatrix} \bar{x}_2 + \bar{x}_2^T \begin{bmatrix} S^2 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix} \bar{x}_2, \end{aligned}$$

with $\bar{\rho}_1 \in [\lambda_1, \lambda_p]$, $\bar{\rho}_2 \in [\lambda_{p+1}, 1]$ and satisfying

$$\bar{x}_1^T M_\gamma \bar{x}_1 = \bar{\rho}_1\|\bar{x}_1\|^2 \quad \text{and} \quad \bar{x}_2^T \tilde{M}_\gamma \bar{x}_2 = \bar{\rho}_2\|\bar{x}_2\|^2.$$

Combining with (6.38) and (6.39), we then have

$$\bar{\nu}^2(\|\bar{x}_1\|^2 + \|\bar{x}_2\|^2) - \bar{\nu}(\bar{\rho}_1\|\bar{x}_1\|^2 + \bar{\rho}_2\|\bar{x}_2\|^2) = \bar{\nu}^2(\|\bar{y}_1\|^2 + \|\bar{y}_2\|^2). \tag{6.41}$$

Observe that $\|\bar{x}_1\|^2 + \|\bar{x}_2\|^2 \neq 0$, since otherwise we would have that $\bar{x}_1 = \bar{x}_2 = 0$ and, by (6.38) and (6.39), that $\bar{y}_1 = \bar{y}_2 = 0$, since $\bar{\nu} > 0$, implying a zero eigenvector $\bar{x}$. From (6.41), we can then deduce, since $\bar{\nu} > 0$, that

$$\begin{aligned} \bar{\nu} &= \frac{\bar{\rho}_1\|\bar{x}_1\|^2 + \bar{\rho}_2\|\bar{x}_2\|^2}{\|\bar{x}_1\|^2 + \|\bar{x}_2\|^2} + \bar{\nu}\frac{\|\bar{y}_1\|^2 + \|\bar{y}_2\|^2}{\|\bar{x}_1\|^2 + \|\bar{x}_2\|^2} \\ &\geq \frac{\bar{\rho}_1\|\bar{x}_1\|^2 + \bar{\rho}_2\|\bar{x}_2\|^2}{\|\bar{x}_1\|^2 + \|\bar{x}_2\|^2} \\ &= \bar{\rho}_1(1-\theta) + \bar{\rho}_2\theta, \end{aligned}$$

where

$$\theta = \frac{\|\bar{x}_2\|^2}{\|\bar{x}_1\|^2 + \|\bar{x}_2\|^2} \in [0,1].$$

Assume that $\theta \geq 1/2$. It then implies, by definition of $\theta$, that $\|\bar{x}_2\|^2 \geq \|\bar{x}_1\|^2$, which in turn leads to

$$\bar{\nu} = \bar{\rho}_1(1-\theta) + \bar{\rho}_2\theta \in \left[\frac{\bar{\rho}_1 + \bar{\rho}_2}{2}, \bar{\rho}_2\right], \tag{6.42}$$

since $\bar{\rho}_1 < \bar{\rho}_2$ by (6.33). One thus has, by (6.42), that

$$\bar{\nu} \geq \frac{\bar{\rho}_1 + \bar{\rho}_2}{2} \geq \frac{\bar{\rho}_2}{2} \geq \gamma/2,$$

since $\bar{\rho}_2 \geq \lambda_{p+1} > \gamma$ by (6.33), which contradicts the assumption that $\bar{\nu} < \gamma/2$. Hence one has that

$$\|\bar{x}_2\|^2 < \|\bar{x}_1\|^2, \tag{6.43}$$

implying that $\bar{x}_1 \neq 0$.

Let now consider $\bar{x}_2$. Assume that $\bar{x}_2 = 0$, by (6.30) and (6.31), we have, on one hand,

$$S\bar{y}_1 = 0, \tag{6.44}$$
$$C\bar{x}_1 = \bar{\nu}\bar{y}_1. \tag{6.45}$$

On the other hand, one has that

$$C\bar{x}_1 = \bar{x}_1. \tag{6.46}$$

Indeed, first observe that by the assumption $c_{\min} := \min_{i=1:p}\{c_i\} > 0$, one has that $c_i > 0$ for all $i = 1, \ldots, p$. If $c_i = 1$, then obviously (6.46) holds for index $i$. If $c_i \neq 1$, implying that $s_i \neq 0$ by (6.7) then the corresponding component in $\bar{y}_1$ is equal to zero, by (6.44), and so does the corresponding component in $\bar{x}_1$, by (6.45), so that again (6.46) is satisfied for this index $i$. Observing that $C^2\bar{x}_1 = \bar{x}_1$ by (6.46) and $\bar{x}_1^T CS = \bar{\nu}\bar{y}_1^T S = 0$ by (6.45) followed by (6.44), we can rewrite (6.36) as

$$-\bar{\nu}^2\|\bar{x}_1\|^2 + \bar{\nu}\bar{\rho}_1\|\bar{x}_1\|^2 + \|\bar{x}_1\|^2 = 0,$$

or, equivalently since $\bar{x}_1 \neq 0$,

$$-\bar{\nu}^2 + \bar{\nu}\bar{\rho}_1 + 1 = 0.$$

The positive root of this last equation in $\bar{\nu}$ gives

$$\bar{\nu} = \frac{\bar{\rho}_1 + \sqrt{\bar{\rho}_1^2 + 4}}{2} > 1,$$

i.e., $\bar{\nu} > \lambda_{\max}(A) = 1$, which leads to a contradiction. Hence, $\bar{x}_2 \neq 0$. □

We now can define the following quantities associated to $0 < \bar{\nu} < \gamma/2$, where $\bar{x}_1$ and $\bar{x}_2$ are nonzero vectors, as guaranteed by Lemma 6.2.

$$\bar{\omega} = \frac{\|\bar{x}_1\|^2}{\|\bar{x}_2\|^2}, \tag{6.47}$$

with $\bar{\omega} > 1$ by (6.43),

$$\bar{\rho}_1 = \frac{\bar{x}_1^T M_\gamma \bar{x}_1}{\|\bar{x}_1\|^2} \in [\lambda_1, \lambda_p], \tag{6.48}$$

and

$$\bar{\rho}_2 = \frac{\bar{x}_2^T \tilde{M}_\gamma \bar{x}_2}{\|\bar{x}_2\|^2} \in [\lambda_{p+1}, 1]. \tag{6.49}$$

Let us also define

$$\bar{\alpha} = \frac{\bar{x}_1^T C^2 \bar{x}_1}{\|\bar{x}_1\|^2}, \tag{6.50}$$

satisfying $0 < c_{\min}^2 \leq \bar{\alpha} \leq 1$,

$$\bar{\beta} = \frac{\bar{x}_2^T \begin{bmatrix} S^2 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix} \bar{x}_2}{\|\bar{x}_2\|^2} \in [0, 1], \tag{6.51}$$

and

$$\bar{\tau} \;=\; \frac{\bar{x}_1^T \begin{bmatrix} CS & 0 & 0 \end{bmatrix} \bar{x}_2}{\|\bar{x}_1\|^2}. \tag{6.52}$$

In the next theorem, we apply the relations introduced in Lemma 6.1 to $\bar{x}$ and $\bar{\nu}$ in order to derive new relations in terms of $\bar{\omega}$, $\bar{\rho}_1$, $\bar{\rho}_2$, $\bar{\alpha}$, $\bar{\beta}$ and $\bar{\tau}$.

**Theorem 6.3** Assume that the matrix (6.27) has an eigenvalue $\bar{\nu}$ satisfying $0 < \bar{\nu} < \gamma/2$, with the associated eigenvector $\bar{x} = \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \bar{y}_1 & \bar{y}_2 \end{bmatrix}^T$ with $\bar{x}_1 \in \mathbb{R}^p$, $\bar{x}_2 \in \mathbb{R}^{n-p}$, $\bar{y}_1 \in \mathbb{R}^p$ and $\bar{y}_2 \in \mathbb{R}^{m-p}$. Let $\bar{\omega}$, $\bar{\rho}_1$, $\bar{\rho}_2$, $\bar{\alpha}$, $\bar{\beta}$ and $\bar{\tau}$ be given by (6.47) to (6.52), respectively and $c_{\min} > 0$. We then have

$$-\bar{\nu}^2 + \bar{\nu}\bar{\rho}_1 + \bar{\alpha} + \bar{\tau} = 0, \tag{6.53}$$
$$-\bar{\nu}^2 + \bar{\nu}\bar{\rho}_2 + \bar{\tau}\bar{\omega} + \bar{\beta} = 0, \tag{6.54}$$
$$\bar{\tau} > -\bar{\alpha}, \tag{6.55}$$
$$\bar{\tau}\bar{\omega} < -\bar{\beta}, \tag{6.56}$$
$$\bar{\tau}^2\bar{\omega} \leq \bar{\alpha}\bar{\beta}. \tag{6.57}$$

**Proof.** We first prove (6.53) and (6.54). Dividing (6.36) and (6.37) by $\|\bar{x}_1\|^2$ and $\|\bar{x}_2\|^2$, respectively, where $\bar{x}_1 \neq 0$ and $\bar{x}_2 \neq 0$ by Lemma 6.2,

$$\bar{\nu}^2 \;=\; \bar{\nu}\bar{\rho}_1 + \frac{\bar{x}_1^T C^2 \bar{x}_1}{\|\bar{x}_1\|^2} + \frac{\bar{x}_1^T \begin{bmatrix} CS & 0 & 0 \end{bmatrix} \bar{x}_2}{\|\bar{x}_1\|^2}$$

$$\bar{\nu}^2 \;=\; \bar{\nu}\bar{\rho}_2 + \frac{\bar{x}_1^T \begin{bmatrix} CS & 0 & 0 \end{bmatrix} \bar{x}_2}{\|\bar{x}_2\|^2} + \frac{\bar{x}_2^T \begin{bmatrix} S^2 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix} \bar{x}_2}{\|\bar{x}_2\|^2}.$$

Using (6.47), (6.50), (6.51) and (6.52), we obtain the desired equalities.

We next prove (6.55). Adding (6.38) to (6.39), it implies

$$0 \leq \bar{\nu}^2(\|\bar{y}_1\|^2 + \|\bar{y}_2\|^2) = \bar{x}_1 C^2 \bar{x}_1 + 2\bar{x}_1^T \begin{bmatrix} CS & 0 & 0 \end{bmatrix} \bar{x}_2 + \bar{x}_2^T \begin{bmatrix} S^2 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix} \bar{x}_2, \tag{6.58}$$

which leads, by (6.47), (6.50), (6.51) and (6.52), that

$$(\bar{\alpha}+\bar{\tau})\|\bar{x}_1\|^2 + (\bar{\tau}\bar{\omega}+\bar{\beta})\|\bar{x}_2\|^2 \geq 0. \tag{6.59}$$

Observing that (6.37) can be written as $(\bar{\tau}\bar{\omega}+\bar{\beta})\|\bar{x}_2\|^2 = \bar{\nu}(\bar{\nu}-\bar{\rho}_2)\|\bar{x}_2\|^2$, (6.59) becomes

$$(\bar{\alpha}+\bar{\tau})\|\bar{x}_1\|^2 + \bar{\nu}(\bar{\nu}-\bar{\rho}_2)\|\bar{x}_2\|^2 \quad \geq \quad 0,$$

or, equivalently, by (6.47),

$$(\bar{\alpha}+\bar{\tau})\bar{\omega} \quad \geq \quad \bar{\nu}(\bar{\rho}_2-\bar{\nu}).$$

As $0 < \bar{\nu} < \gamma/2 < \bar{\rho}_2$ and $\bar{\omega} > 0$, we deduce that $\bar{\alpha}+\bar{\tau} > 0$, which proves (6.55).

We prove (6.56) by contradiction. Assume that $\bar{\tau}\bar{\omega} \geq -\bar{\beta}$, then by (6.54), one has that

$$\bar{\nu}^2 - \bar{\nu}\bar{\rho}_2 = \bar{\nu}(\bar{\nu}-\bar{\rho}_2) \geq 0,$$

which implies, since $\bar{\nu} > 0$, that $\bar{\nu} \geq \bar{\rho}_2 \geq \gamma/2$ and contradicts the assumption $\bar{\nu} < \gamma/2$.

We finally prove (6.57). Multiplying (6.31) by $\bar{x}_1^T C$ implies

$$\bar{x}_1^T C^2 \bar{x}_1 + \bar{x}_1^T C \begin{bmatrix} S & 0 & 0 \end{bmatrix} \bar{x}_2 = (\bar{\alpha}+\bar{\tau})\|\bar{x}_1\|^2 \quad = \quad \bar{\nu}\bar{x}_1^T C \bar{y}_1.$$

Combining this last equality with (6.55) and the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} 0 < (\bar{\alpha}+\bar{\tau})\|\bar{x}_1\|^2 &= \bar{\nu}\bar{x}_1^T C\bar{y}_1 \\ &\leq \bar{\nu}\|C\bar{x}_1\|\|\bar{y}_1\| \\ &= \bar{\nu}\sqrt{\bar{\alpha}}\|\bar{x}_1\|\|\bar{y}_1\|, \end{aligned} \tag{6.60}$$

where the last equality derives from the definition of $\bar{\alpha}$ in (6.50). We can rewrite (6.58) as

$$\bar{\nu}^2(\|\bar{y}_1\|^2 + \|\bar{y}_2\|^2) = (\bar{\alpha}+2\bar{\tau})\|\bar{x}_1\|^2 + \bar{\beta}\|\bar{x}_2\|^2. \tag{6.61}$$

Squaring both sides of (6.60) implies

$$\begin{aligned} (\bar{\alpha}+\bar{\tau})^2\|\bar{x}_1\|^2 &\leq \bar{\nu}^2\bar{\alpha}\|\bar{y}_1\|^2 \\ &\leq \bar{\nu}^2\bar{\alpha}(\|\bar{y}_1\|^2 + \|\bar{y}_2\|^2). \end{aligned}$$

Combining now this last inequality with (6.61) and dividing by $\|\bar{x}_2\|^2$ gives

$$(\bar{\alpha}+\bar{\tau})^2\bar{\omega} \leq (\bar{\alpha}^2 + 2\bar{\alpha}\bar{\tau})\bar{\omega} + \bar{\alpha}\bar{\beta},$$

which, after simplification, yields the desired result (6.57). □

### 6.3.3 Refining the positive lower bound

In the previous section, we have assumed the existence of an eigenvalue $\bar{\nu}$ of $\mathcal{A}_{KKT}$ given by (6.20) satisfying $0 < \bar{\nu} < \gamma/2$ and shown, under the assumption that $c_{\min} = \min_{i=1:p}\{c_i\} > 0$, that $\bar{\nu}$ and its associated eigenvector $\bar{x}$ satisfy the relations (6.53)-(6.57). In order to refine the positive lower bound in (6.21) as given by Rusten and Winther (1992), we proceed in two steps, building two optimization problems successively, whose optimal solution will provide the desired refined positive lower bound.

To build the feasible domain of the first of these two optimization problems, we relax the relations (6.53)-(6.57) by relaxing the quantities $\bar{\nu}$, $\bar{\tau}$ and $\bar{\omega}$ in these relations, which now become the variables $\nu$, $\tau$ and $\omega$ verifying

$$\bar{\rho}_1 \leq \nu \leq \frac{\bar{\rho}_2}{2} \quad \text{and} \quad 1 \leq \omega \leq \omega_{\max}, \tag{6.62}$$

where $\omega_{\max}$ is an upper bound satisfying $\omega_{\max} \geq \bar{\omega}$. The constraints of this optimization problem, let call it $P(\bar{\rho}_1, \bar{\rho}_2, \bar{\alpha}, \bar{\beta}, \omega_{\max})$, can thus be written as

$$\mathcal{C}(P) \equiv \begin{cases} -\nu^2 + \nu\bar{\rho}_1 + \bar{\alpha} + \tau = 0, & \text{(6.63a)} \\ -\nu^2 + \nu\bar{\rho}_2 + \bar{\beta} + \tau\omega = 0, & \text{(6.63b)} \\ \bar{\rho}_1 \leq \nu \leq \dfrac{\bar{\rho}_2}{2}, & \text{(6.63c)} \\ 1 \leq \omega \leq \omega_{\max}, & \text{(6.63d)} \\ \tau \geq -\bar{\alpha}, & \text{(6.63e)} \\ \tau\omega \leq -\bar{\beta}, & \text{(6.63f)} \\ \tau^2\omega \leq \bar{\alpha}\bar{\beta}. & \text{(6.63g)} \end{cases}$$

We then minimize $\nu$ over the set $(\nu, \tau, \omega)$ satisfying these constraints, i.e., we consider the following optimization problem

$$P(\bar{\rho}_1, \bar{\rho}_2, \bar{\alpha}, \bar{\beta}, \omega_{\max}) = \min_{(\nu,\tau,\omega)\in\mathcal{C}(P)} \nu. \tag{6.64}$$

Note that $\mathcal{C}(P)$ is nonempty, since $(\bar{\nu}, \bar{\tau}, \bar{\omega}) \in \mathcal{C}(P)$, and that $\nu$ no more represents an eigenvalue of the matrix $\mathcal{A}_{KKT}$ in this problem. Instead the optimal value $\nu_0$ of (6.64), whose existence is guaranteed by the compactness of the feasible set $\mathcal{C}(P)$ (see, e.g., the Weierstrass Theorem in Hiriart-Urruty, 1996), gives a lower bound on $\bar{\nu}$ (since $(\bar{\nu}, \bar{\tau}, \bar{\omega}) \in \mathcal{C}(P)$). Observe that $\tau$ in (6.64) satisfies $\tau \leq 0$ by (6.63f), since $\omega \geq 1$ and $\bar{\beta} \in [0, 1]$, by (6.51). Finally, we have the necessary condition

$$c_{\min} < 1, \tag{6.65}$$

since otherwise $c_i = 1\ \forall i = 1, \ldots, p$ and thus $s_i = 0\ \forall i = 1, \ldots, p$ by (6.7) implying that $\bar{\alpha} = 1$ and $\bar{\tau} = 0$ by (6.50) and (6.52), respectively. Equation (6.54) becomes $-\bar{\nu}^2 + \bar{\nu}\bar{\rho}_2 + \bar{\beta} = 0$ which implies that the positive solution satisfies $\bar{\nu} = \frac{\rho_2 + \sqrt{\rho_2^2 + 4\bar{\beta}}}{2} \geq \bar{\rho}_2$ since $\bar{\beta} \geq 0$, which contradicts (6.62).

Before studying problem (6.64), we establish a bound on the positive solution $\nu$ of equations (6.63a) and (6.63b) that will prove to be useful.

**Lemma 6.4** Assume that $\bar{\alpha} > 0$, $\bar{\beta} \geq 0$, $\omega \geq 1$ and $\tau \leq 0$. Then the system of equations (6.63a) and (6.63b) in $\nu$ has a unique positive solution satisfying

$$\nu \geq \frac{\omega\bar{\rho}_1 + \bar{\rho}_2 + \sqrt{\tilde{\Delta}}}{2(\omega+1)}, \tag{6.66}$$

where $\tilde{\Delta} = (\omega\bar{\rho}_1 + \bar{\rho}_2)^2 + 4(\omega+1)\left(\sqrt{\bar{\alpha}\omega} - \sqrt{\bar{\beta}}\right)^2$.

**Proof.** Multiplying (6.63a) by $\omega$ and adding it to (6.63b), we get the equation

$$(\omega+1)\nu^2 - (\omega\bar{\rho}_1 + \bar{\rho}_2)\nu - \big(\omega(\bar{\alpha}+2\tau) + \bar{\beta}\big) = 0, \tag{6.67}$$

whose roots are given by

$$\nu_{1,2} \quad = \quad \frac{\omega\bar{\rho}_1 + \bar{\rho}_2 \pm \sqrt{\Delta}}{2(\omega+1)},$$

where $\Delta = (\omega\bar{\rho}_1 + \bar{\rho}_2)^2 + 4(\omega+1)(\omega\bar{\alpha} + 2\omega\tau + \bar{\beta})$. By (6.63g) together with $\omega > 0$, we have that

$$\tau^2\omega^2 \leq \bar{\alpha}\bar{\beta}\omega,$$

or, equivalently, since $\tau \leq 0$, $\bar{\alpha} > 0$ and $\bar{\beta} \geq 0$,

$$\tau\omega \geq -\sqrt{\bar{\alpha}\bar{\beta}\omega}.$$

This inequality implies that

$$\begin{aligned} \omega\bar{\alpha} + 2\omega\tau + \bar{\beta} \quad &\geq \quad \omega\bar{\alpha} - 2\sqrt{\bar{\alpha}\bar{\beta}\omega} + \bar{\beta} \\ &= \quad \left(\sqrt{\bar{\alpha}\omega} - \sqrt{\bar{\beta}}\right)^2 \\ &\geq \quad 0, \end{aligned}$$

so that, on one hand, $\Delta \geq (\omega\bar{\rho}_1 + \bar{\rho}_2)^2$ for $\omega \geq 1$, yielding $\nu_1 \leq 0$. On the other hand, it implies that $\Delta \geq \tilde{\Delta}$, so that the unique positive solution of (6.63a) and (6.63b) is $\nu_2$ and satisfies (6.66). □

Our study of an optimal solution of the optimization problem (6.64) starts by identifying the constraints which are potentially active at optimality.

**Theorem 6.5** Consider problem $P(\bar{\rho}_1, \bar{\rho}_2, \bar{\alpha}, \bar{\beta}, \omega_{\max})$ defined in (6.64) where $\omega_{\max} \geq \bar{\omega}$ and $\bar{\omega}, \bar{\rho}_1, \bar{\rho}_2, \bar{\alpha}$ and $\bar{\beta}$ are given by (6.47)-(6.51). Then the constraints in (6.64) possibly active at a global solution are,

$$\begin{aligned} \omega &\leq \omega_{\max} \\ \tau^2\omega &\leq \bar{\alpha}\bar{\beta}. \end{aligned}$$

**Proof.**

1. First let us prove that the lower bound in constraint (6.63d) ($\omega = 1$) is not active. By Lemma 6.4 when $\omega = 1$, we have that

$$\nu \geq \frac{\bar{\rho}_1 + \bar{\rho}_2 + \sqrt{(\bar{\rho}_1 + \bar{\rho}_2)^2 + 8(\sqrt{\bar{\alpha}} + \sqrt{\bar{\beta}})^2}}{4} \geq \frac{(\bar{\rho}_1 + \bar{\rho}_2)}{2} > \frac{\bar{\rho}_2}{2},$$

since $\bar{\rho}_1 > 0$. This is incompatible with (6.63c).

2. Let us prove that the constraint (6.63e) is not active. Assuming that $\tau = -\bar{\alpha}$, we have by (6.63a) and (6.63b),

$$\begin{aligned} -\nu^2 - \nu\bar{\rho}_1 &= 0, & (6.68) \\ -\nu^2 + \nu\bar{\rho}_2 + \bar{\beta} - \bar{\alpha}\omega &= 0. & (6.69) \end{aligned}$$

Equation (6.68) implies that either $\nu = 0$ or $\nu = \bar{\rho}_1$. Since $\nu \geq \bar{\rho}_1 > 0$ by (6.63c), we have that the unique solution of (6.68) is $\nu = \bar{\rho}_1$, and it follows that (6.69) becomes

$$-(\bar{\rho}_1)^2 + \bar{\rho}_1\bar{\rho}_2 + \bar{\beta} - \bar{\alpha}\omega = 0,$$

or, equivalently,

$$\bar{\alpha}\omega = \bar{\rho}_1(\bar{\rho}_2 - \bar{\rho}_1) + \bar{\beta}. \quad (6.70)$$

Multiplying (6.70) by $\bar{\alpha}$, we obtain

$$\bar{\alpha}^2\omega = \tau^2\omega = \bar{\alpha}(\bar{\rho}_1(\bar{\rho}_2 - \bar{\rho}_1) + \bar{\beta}).$$

Since $\bar{\alpha} > 0$, $\bar{\rho}_1 > 0$ and $\bar{\rho}_2 - \bar{\rho}_1 > 0$, one deduces that $\tau^2\omega > \bar{\alpha}\bar{\beta}$, which contradicts (6.63g).

3. Now, we prove that (6.63f) is not active. By contradiction, let assume that $\tau\omega+\bar{\beta}=0$. By (6.63b), we have that $-\nu^2+\nu\bar{\rho}_2=0$, or equivalently, $\nu=0$ or $\nu=\bar{\rho}_2$, which is impossible by (6.63c).

4. It remains to prove that both bounds of (6.63c) are inactive at a global solution. First $\nu=\bar{\rho}_1$ implies by (6.63a) that (6.63e) is active, which is impossible by the second step of the proof. If $\nu=\bar{\rho}_2/2$, then it is not a global solution since $\bar{\nu}<\gamma/2<\bar{\rho}_2/2=\nu$ provides a lower objective function value and $(\bar{\nu},\bar{\tau},\bar{\omega})\in\mathcal{C}(P)$).

□

Assuming that $\omega=\omega_{\max}$ (i.e., the upper bound of (6.63d) is active) at a global solution of problem $P(\bar{\rho}_1,\bar{\rho}_2,\bar{\alpha},\bar{\beta},\omega_{\max})$ given by (6.64), we now derive a lower bound on the optimal value of problem (6.64), hence yielding a first potentially refined positive lower bound on the positive eigenvalues of $\mathcal{A}_{KKT}$.

**Theorem 6.6** Consider problem $P(\bar{\rho}_1,\bar{\rho}_2,\bar{\alpha},\bar{\beta},\omega_{\max})$ defined in (6.64) where $\omega_{\max}\geq\bar{\omega}$ and $\bar{\omega},\bar{\rho}_1,\bar{\rho}_2,\bar{\alpha}$ and $\bar{\beta}$ are given by (6.47)-(6.51). If $\omega=\omega_{\max}$ at a global solution then

$$\nu_0\geq\frac{1}{2}\left(\bar{\rho}_1+\sqrt{\bar{\rho}_1^2+4c_{\min}^2}\right). \tag{6.71}$$

**Proof.** By Lemma 6.4 with $\omega=\omega_{\max}$, we have

$$\nu_0\geq\frac{1}{2}\left(\frac{\omega_{\max}\bar{\rho}_1+\bar{\rho}_2+\sqrt{\tilde{\Delta}}}{\omega_{\max}+1}\right), \tag{6.72}$$

where $\tilde{\Delta}=(\omega_{\max}\bar{\rho}_1+\bar{\rho}_2)^2+4(\omega_{\max}+1)\left(\sqrt{\bar{\alpha}\omega_{\max}}-\sqrt{\bar{\beta}}\right)^2$. Defining the quantities

$$\tilde{\rho}=\frac{\omega_{\max}\bar{\rho}_1+\bar{\rho}_2}{\omega_{\max}+1}$$

and

$$\tilde{\alpha}=\left(\sqrt{\bar{\alpha}}\sqrt{\frac{\omega_{\max}}{\omega_{\max}+1}}-\sqrt{\frac{\bar{\beta}}{\omega_{\max}+1}}\right),$$

we can rewrite (6.72) as

$$\nu_0\geq\frac{1}{2}\left(\tilde{\rho}+\sqrt{\tilde{\rho}^2+4\tilde{\alpha}^2}\right).$$

Observing that $\omega_{\max}$ can be taken as large as we want, provided $\omega_{\max} \geq \bar{\omega}$, and that

$$\lim_{\omega_{\max}\to\infty} \tilde{\rho} = \bar{\rho}_1 \quad \text{and} \quad \lim_{\omega_{\max}\to\infty} \tilde{\alpha} = \sqrt{\bar{\alpha}},$$

one can conclude, using the bound $\alpha \geq c_{\min}^2$, that

$$\nu_0 \geq \frac{1}{2}\left(\bar{\rho}_1 + \sqrt{\bar{\rho}_1^2 + 4c_{\min}^2}\right).$$

□

We next assume that $\omega < \omega_{\max}$ and apply the first-order necessary optimality conditions given by F. John Theorem 1.5 in Chapter 1 to problem (6.64): there exist $t$, $u$, $v$, $p \in \mathbb{R}$ not all equal to zero such that $p \geq 0$,

$$\begin{bmatrix}1\\0\\0\end{bmatrix} t - \begin{bmatrix}-2\nu+\bar{\rho}_1\\1\\0\end{bmatrix} u - \begin{bmatrix}-2\nu+\bar{\rho}_2\\\omega\\\tau\end{bmatrix} v - \begin{bmatrix}0\\-2\tau\omega\\-\tau^2\end{bmatrix} p = \begin{bmatrix}0\\0\\0\end{bmatrix}, \tag{6.73}$$

and

$$p(\bar{\alpha}\bar{\beta} - \tau^2\omega) = 0. \tag{6.74}$$

Note first that $\tau \neq 0$. Indeed, otherwize $\bar{\beta} \leq 0$ by (6.63f), implying that $\bar{\beta} = 0$ since $\bar{\beta} \in [0, 1]$, which in turn would give, by (6.63b),

$$-\nu^2 + \nu\bar{\rho}_2 = \nu(\bar{\rho}_2 - \nu) = 0,$$

so that $\nu = 0$ or $\nu = \bar{\rho}_2$, in contradiction with (6.63c). We next obtain that $p \neq 0$ since otherwise $\tau v = 0$ by the third equality in (6.73), and thus $v = 0$ which in turn implies $u = 0$ by the second equality in (6.73) followed by $t = 0$ by the first equality in (6.73). This is incompatible with the assumption that $t, u, v$ and $p$ can not be all equal to zero. The complementarity condition (6.74) then ensures that the constraint (6.63g) is active, i.e., $\tau^2\omega = \bar{\alpha}\bar{\beta}$, so that, since $\tau < 0$ by (6.63e) and $\omega \neq 0$ by (6.63d), we can set

$$\tau = -\sqrt{\frac{\bar{\alpha}\bar{\beta}}{\omega}}.$$

Let us now denote by $\tau_0$ and $\omega_0$ the associated quantities to $\nu_0$, a global solution of problem $P(\bar{\rho}_1, \bar{\rho}_2, \bar{\alpha}, \bar{\beta}, \omega_{\max})$. We then deduce that $\tau_0 = -\sqrt{\frac{\bar{\alpha}\bar{\beta}}{\omega_0}}$. By setting

$$\delta_0 = \sqrt{\frac{\bar{\alpha}\omega_0}{\bar{\beta}}},$$

we can rewrite (6.63a) and (6.63b), observing that $\tau_0 = -\frac{\bar{\alpha}}{\delta_0} = \frac{-\delta_0\bar{\beta}}{\omega_0}$,

$$-\nu_0^2 + \nu_0\bar{\rho}_1 + \bar{\alpha}\left(1 + \frac{\tau_0}{\bar{\alpha}}\right) = -\nu_0^2 + \nu_0\bar{\rho}_1 + \bar{\alpha}\left(1 - \frac{1}{\delta_0}\right) = 0, \quad (6.75)$$

$$-\nu_0^2 + \nu_0\bar{\rho}_2 + \bar{\beta}\left(1 + \frac{\tau_0\omega_0}{\bar{\beta}}\right) = -\nu_0^2 + \nu_0\bar{\rho}_2 + \bar{\beta}\,(1 - \delta_0) = 0. \quad (6.76)$$

We also have by (6.63e) (which is inactive at a solution), that

$$\tau_0 = -\frac{\bar{\alpha}}{\delta_0} > -\bar{\alpha},$$

so that $\delta_0 > 1$.

In order to proceed in our search for a refined positive lower bound in (6.21), we now build a second constrained optimization problem in the same spirit, i.e., with the aim to derive a lower bound on $\nu_0$ (and thus on $\bar{\nu}$ since $\nu_0 \leq \bar{\nu}$). To that, we relax the quantities $\nu_0$, $\delta_0$, $\bar{\alpha}$ and $\bar{\beta}$ and consider the optimization problem

$$\tilde{P}(\bar{\rho}_1, \bar{\rho}_2) \equiv \min_{(\nu,\delta,\alpha,\beta)\in\mathcal{C}(\tilde{P})} \quad \nu \quad (6.77)$$

where the feasible set $\mathcal{C}(\tilde{P})$ is defined by

$$\mathcal{C}(\tilde{P}) \equiv \begin{cases} -\nu^2 + \nu\bar{\rho}_1 + \alpha\left(1 - \frac{1}{\delta}\right) = 0, & (6.78a) \\ -\nu^2 + \nu\bar{\rho}_2 + \beta\,(1 - \delta) = 0, & (6.78b) \\ \bar{\rho}_1 \leq \nu \leq \frac{\bar{\rho}_2}{2}, & (6.78c) \\ 1 \leq \delta \leq \delta_{\max}, & (6.78d) \\ c_{\min}^2 \leq \alpha \leq 1, & (6.78e) \\ 0 \leq \beta \leq 1, & (6.78f) \end{cases}$$

where $\delta_{\max}$ is an upper bound satisfying $\delta_{\max} \geq \delta_0 > 1$ and where $c_{\min} < 1$ (from the necessary condition (6.65)). Note that $\mathcal{C}(\tilde{P})$ is nonempty since $(\nu_0, \delta_0, \bar{\alpha}, \bar{\beta}) \in \mathcal{C}(\tilde{P})$ by (6.75), (6.76), (6.63c) satisfied by $\nu_0$, (6.50) and (6.51). Again, the compactness of the feasible set $\mathcal{C}(\tilde{P})$ guarantees the existence of an optimal value $\nu_{\inf}$ for problem $\tilde{P}(\bar{\rho}_1, \bar{\rho}_2)$ with $\nu_{\inf} \leq \nu_0$.

The next theorem identifies the constraints of $\tilde{P}(\bar{\rho}_1, \bar{\rho}_2)$ which are potentially active at optimality.

**Theorem 6.7** Consider problem $\tilde{P}(\bar{\rho}_1, \bar{\rho}_2)$ defined in (6.77) where $\delta_{\max} \geq \delta_0$, and $\bar{\rho}_1$ and $\bar{\rho}_2$ are given by (6.48) and (6.49), respectively. Then, the constraints in (6.77) possibly active at a global solution are,

$$\begin{array}{rcl} \delta & \leq & \delta_{\max} \\ c_{\min}^2 \leq & \alpha & \leq 1 \\ \beta & \leq & 1 \end{array}$$

**Proof.**

1. Let first show that the lower bound in (6.78d) ($\delta = 1$) is not active at optimality. Indeed, if $\delta = 1$, we get by (6.78b),

$$-\nu^2 + \nu\bar{\rho}_2 = \nu\,(\bar{\rho}_2 - \nu) = 0,$$

   so that $\nu = 0$ or $\nu = \bar{\rho}_2$, in contradiction with (6.78c).

2. Using the same argument, we have that the lower bound in (6.78f) ($\beta = 0$) is not active.

3. It remains to prove that both bounds of (6.78c) are inactive at a global solution. First $\nu = \bar{\rho}_1$ implies by (6.78a) that $\alpha(1 - \frac{1}{\delta}) = 0$, which is impossible since $\alpha \neq 0$ and $\delta \neq 1$. If $\nu = \frac{\bar{\rho}_2}{2}$, then it is not a global solution since $\nu_0 \leq \bar{\nu} < \frac{\gamma}{2} < \frac{\bar{\rho}_2}{2} = \nu$ provides a lower objective function value and $(\nu_0, \delta_0, \bar{\alpha}, \bar{\beta} \in \mathcal{C}(\tilde{P}))$.

□

Similarly to the way we have proceeded for problem (6.64), we first consider the case where $\delta = \delta_{\max}$ (i.e., the upper bound of (6.78d) is active) at a global solution of problem $\tilde{P}(\bar{\rho}_1, \bar{\rho}_2)$. As shown in the next theorem, we retrieve the same lower bound on the positive eigenvalues of $\mathcal{A}_{KKT}$ as given in Theorem 6.6.

**Theorem 6.8** Consider problem $\tilde{P}(\bar{\rho}_1, \bar{\rho}_2)$ defined in (6.77) where $\delta_{\max} \geq \delta_0$, and $\bar{\rho}_1$ and $\bar{\rho}_2$ are given by (6.48) and (6.49), respectively. If $\delta = \delta_{\max}$ at a global solution then

$$\nu_{\inf} \geq \frac{1}{2}\left(\bar{\rho}_1 + \sqrt{\bar{\rho}_1 + 4c_{\min}^2}\right).$$

**Proof.** By (6.78a) with $\delta = \delta_{\max}$, we obtain

$$-\nu^2 + \nu\bar{\rho}_1 + \alpha\left(1 - \frac{1}{\delta_{\max}}\right) = 0$$

whose roots are given by

$$\nu_{\inf} = \frac{1}{2}\left(\bar{\rho}_1 \pm \sqrt{\bar{\rho}_1^2 + 4\alpha\left(1 - \frac{1}{\delta_{\max}}\right)}\right),$$

where $\bar{\rho}_1^2 + 4\alpha\left(1 - \frac{1}{\delta_{\max}}\right) > 0$ as $\alpha > 0$ and $\delta_{\max} \geq \delta_0 > 1$. Excluding the negative root, one has that

$$\nu = \frac{1}{2}\left(\bar{\rho}_1 + \sqrt{\bar{\rho}_1^2 + 4\alpha\left(1 - \frac{1}{\delta_{\max}}\right)}\right) > \bar{\rho}_1.$$

Observing that $\delta_{\max}$ in problem $\tilde{P}(\bar{\rho}_1, \bar{\rho}_2)$ can be taken as large as we want, provided $\delta_{\max} \geq \delta_0$, and that

$$\lim_{\delta_{\max}\to\infty} \frac{1}{2}\left(\bar{\rho}_1 + \sqrt{\bar{\rho}_1^2 + 4\alpha\left(1 - \frac{1}{\delta_{\max}}\right)}\right) = \frac{1}{2}\left(\bar{\rho}_1 + \sqrt{\bar{\rho}_1^2 + 4\alpha}\right),$$

one can conclude, using the bound $\alpha \geq c_{\min}^2$, that

$$\nu_{\inf} \geq \frac{1}{2}\left(\bar{\rho}_1 + \sqrt{\bar{\rho}_1^2 + 4c_{\min}^2}\right).$$

□

We next, and finally, consider the case where $\delta < \delta_{\max}$. Using again F. John Theorem 1.5 (see Chapter 1), we have that there exist $t$, $u$, $v$, $p$, $q$, $r \in \mathbb{R}$ not all equal to zero such that $p \geq 0$, $q \geq 0$, $r \geq 0$,

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} t - \begin{bmatrix} -2\nu + \bar{\rho}_1 \\ \frac{\alpha}{\delta^2} \\ 1 - \frac{1}{\delta} \\ 0 \end{bmatrix} u - \begin{bmatrix} -2\nu + \bar{\rho}_2 \\ -\beta \\ 0 \\ 1 - \delta \end{bmatrix} v - \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} p - \begin{bmatrix} 0 \\ 0 \\ -1 \\ 0 \end{bmatrix} q + \begin{bmatrix} 0 \\ 0 \\ 0 \\ -1 \end{bmatrix} r = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \tag{6.79}$$

and

$$\begin{aligned} p(c_{\min}^2 - \alpha) &= 0, && (6.80)\\ q(\alpha - 1) &= 0, && (6.81)\\ r(\beta - 1) &= 0, && (6.82) \end{aligned}$$

If $r = 0$, then $v = 0$ by the last equality in (6.79) since $\delta > 1$, and consequently $u = 0$ by the second equality in (6.79) and since $\alpha/\delta^2 \neq 0$. The first and third equalities of (6.79) then imply $t = 0$ and $p = q$, respectively. Since $t$, $u$, $v$, $p$, $q$, $r$ cannot be all equal to zero, then one must have $p = q \neq 0$, which implies, by the complementarity conditions (6.80) and (6.81), that $\alpha = 1 = c_{\min}^2$, which is impossible. Assume now that $r > 0$. Then $\beta = 1$ by (6.82), and the last equality in (6.79) yields $(1 - \delta)v = r$. This, together with $r > 0$ and $\delta > 1$, implies that $v < 0$. The second equality in (6.79) then gives

$$-\frac{\alpha}{\delta^2}u + v = 0,$$

implying that $u < 0$ since $\alpha > 0$. By the third equality in (6.79), we get

$$-\left(1 - \frac{1}{\delta}\right)u - p + q = 0,$$

that is, $p - q > 0$ since $\delta > 1$. As (6.80) and (6.81) with $c_{\min}^2 < 1$ imply that either $p$ or $q$ must be zero, the only possibility is to have $q = 0$ and $p > 0$ (since otherwise $p = 0$ and $q < 0$, in contradiction with the sign condition $q \geq 0$ on the multiplier $q$). We thus have $\alpha = c_{\min}^2$ by (6.80).

Rewriting $\tilde{P}(\bar{\rho}_1, \bar{\rho}_2)$ with $\beta = 1$ and $\alpha = c_{\min}^2$, we get

$$\begin{array}{ll} \min\limits_{(\nu,\delta)} & \nu \\ \text{s.t.} & \left\{ \begin{array}{c} -\nu^2 + \nu\bar{\rho}_1 + c_{\min}^2(1 - \frac{1}{\delta}) = 0, \\ -\nu^2 + \nu\bar{\rho}_2 + (1 - \delta) = 0, \\ \bar{\rho}_1 \leq \nu \leq \frac{\bar{\rho}_2}{2}, \\ 1 \leq \delta \leq \delta_{\max}, \end{array} \right. \end{array} \tag{6.83}$$

from which we can deduce another last lower bound on the positive eigenvalues of $\mathcal{A}_{KKT}$.

**Theorem 6.9** Consider problem $\tilde{P}(\bar{\rho}_1, \bar{\rho}_2)$ defined in (6.83) where $\delta_{\max} \geq \delta_0$, and $\bar{\rho}_1$ and $\bar{\rho}_2$ are given by (6.48) and (6.49), respectively. Then the optimal value satisfies

$$\nu_{\inf} \geq \frac{\bar{\rho}_1 + \frac{4}{5}c_{\min}^2\bar{\rho}_2}{1 + \frac{4}{5}c_{\min}^2}.$$

**Proof.** Multiplying the first equation in (6.83) by $\delta$, the second by $c_{\min}^2$ and summing, we have

$$-\nu^2(\delta + c_{\min}^2) + \nu(\bar{\rho}_1\delta + \bar{\rho}_2 c_{\min}^2) = 0,$$

this last equation has a single feasible solution that we can express as a function $\delta$,

$$\nu(\delta) = \frac{\bar{\rho}_1 \delta + \bar{\rho}_2 c_{\min}^2}{\delta + c_{\min}^2}.$$

Observing that

$$\begin{aligned} \nu'(\delta) &= \frac{\bar{\rho}_1(\delta + c_{\min}^2) - (\bar{\rho}_1 \delta + \bar{\rho}_2 c_{\min}^2)}{(\delta + c_{\min}^2)^2} \\ &= \frac{(\bar{\rho}_1 - \bar{\rho}_2) c_{\min}^2}{(\delta + c_{\min}^2)^2} < 0, \end{aligned}$$

since $\bar{\rho}_1 < \bar{\rho}_2$, we have that $\nu(\delta)$ is a strictly decreasing function. On the other hand, the second equation in (6.83) requires, to have a solution, that

$$\bar{\rho}_2^2 + 4(1 - \delta) \geq 0,$$

that is, the largest possible value for $\delta$ at optimality is $\frac{\bar{\rho}_2^2 + 4}{4}$. We can thus conclude, since $\bar{\rho}_2 \leq 1$ by (6.49), that

$$\nu_{\inf} = \nu\left(\frac{\bar{\rho}_2^2 + 4}{4}\right) \geq \nu(5/4) = \frac{\bar{\rho}_1 + \frac{4}{5} c_{\min}^2 \bar{\rho}_2}{1 + \frac{4}{5} c_{\min}^2}. \tag{6.84}$$

□

Gathering the results from Theorem 6.6, 6.8 and 6.9, we can now formulate our final result

**Theorem 6.10** Assume that the matrix (6.27) has an eigenvalue $\bar{\nu}$ satisfying $0 < \bar{\nu} < \gamma/2$ and let $C \in \mathbb{R}^{p \times p}$ given by (6.23) be such that $0 < c_{\min} = \min_{i=1:p}\{c_i\}$. Then the eigenvalues of $\mathcal{A}_{KKT}$ are bounded within

$$\left[\frac{\lambda_{\min}(A) - \sqrt{\lambda_{\min}^2(A) + 4}}{2}, \frac{1 - \sqrt{5}}{2}\right] \bigcup \left[b_{\inf}, \frac{1 + \sqrt{5}}{2}\right], \tag{6.85}$$

where $b_{\inf} = \min\left(\frac{1}{2}\left(\bar{\rho}_1 + \sqrt{\bar{\rho}_1^2 + 4c_{\min}^2}\right), \frac{\bar{\rho}_1 + \frac{4}{5} c_{\min}^2 \bar{\rho}_2}{1 + \frac{4}{5} c_{\min}^2}\right)$.

Note first that

$$\frac{\bar{\rho}_1 + \frac{4}{5} c_{\min}^2 \bar{\rho}_2}{1 + \frac{4}{5} c_{\min}^2} > \bar{\rho}_1.$$

Indeed, we have that

$$\begin{aligned}\frac{\bar{\rho}_1 + \frac{4}{5}c_{\min}^2\bar{\rho}_2}{1+\frac{4}{5}c_{\min}^2} - \bar{\rho}_1 &= \frac{\bar{\rho}_1 + \frac{4}{5}c_{\min}^2\bar{\rho}_2 - \bar{\rho}_1\left(1+\frac{4}{5}c_{\min}^2\right)}{1+\frac{4}{5}c_{\min}^2} \\ &= \frac{\frac{4}{5}c_{\min}^2\left(\bar{\rho}_2 - \bar{\rho}_1\right)}{1+\frac{4}{5}c_{\min}^2} \\ &> 0,\end{aligned}$$

since $\bar{\rho}_1 < \bar{\rho}_2$ and $c_{\min} > 0$. Furthermore, we observe that

$$\frac{1}{2}\left(\bar{\rho}_1 + \sqrt{\bar{\rho}_1^2 + 4c_{\min}^2}\right) > \bar{\rho}_1,$$

since $c_{\min} > 0$. We conclude that $b_{\inf} > \bar{\rho}_1 \geq \lambda_{\min}(A)$ implying that the lower bound of the right interval in (6.21) is refined. We can also observe that when $c_{\min} \to 0$, we have that $b_{\inf} \to \bar{\rho}_1$ with $\bar{\rho}_1 \geq \lambda_{\min}(A)$, which is consistent with the result given in Rusten and Winther (1992). For instance, we set $\lambda_{\min}(A) = 10^{-8}$ and $\lambda_{p+1} = \gamma$ with $\gamma = \lambda_{\max}(A)/10 = 10^{-1}$ since we have assumed that a first level of preconditioning has been applied so that $\lambda_{\max}(A) = 1$. We choose $\bar{\rho}_1 = 10^{-8}$ and $\bar{\rho}_2 = 10^{-1}$ and in Figure 6.3, we illustrate the behaviour of functions

$$b_1(c_{\min}) := \frac{1}{2}\left(10^{-8} + \sqrt{10^{-16} + 4c_{\min}^2}\right),$$

and

$$b_2(c_{\min}) := \frac{10^{-8} + \frac{4}{5}c_{\min}^2 10^{-1}}{1+\frac{4}{5}c_{\min}^2},$$

such that

$$b_{\inf} = \min\left(b_1(c_{\min}), b_2(c_{\min})\right)$$

for $c_{\min} \in [0,1]$. We observe that $b_2(c_{\min})$ is below $b_1(c_{\min})$ for all values of $c_{\min}$. By Rusten and Winther (1992), the lower bound on the positive eigenvalues of $\mathcal{A}_{KKT}$ is given by $\lambda_{\min}(A) = 10^{-8}$ and we can see in Figure 6.3 that if $c_{\min} = 10^{-1}$, the lower bound becomes $7.9\ 10^{-4}$.

## 6.4 Reduced spectral information

The natural idea that arises from these considerations in previous Sections is to incorporate into the approximation of the Schur complement inverse (6.9) not all the invariant subspace $\mathcal{I}m(U_\gamma)$, but only those principal vectors associated to cosines of the principal angles less than $\tau\sqrt{\gamma/\alpha}$ (with a value of $\tau$ within $[0.5, 2]$, for instance) as studied in inequality (6.12). This may enable us to reduce substantially the size of the low rank update

$$\frac{1}{\alpha}C_\gamma^{-1}V_\gamma^T\Lambda_\gamma V_\gamma C_\gamma^{-1} + I_p$$

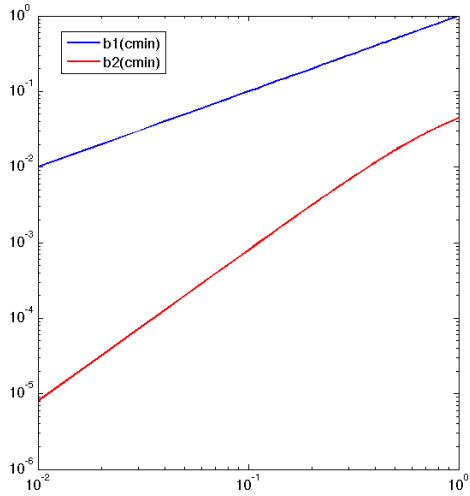



Figure 6.3 – $b_1(c_{\min})$ and $b_2(c_{\min})$ for $c_{\min} \in [0, 1]$.

in $S_\gamma^{-1}$ while keeping the value of $\gamma$ large enough to improve even further the speed of convergence in Minres.

Let us denote by $\ell \ll p$ the number of cosines less than the above threshold, and by $V_\ell \in \mathbb{R}^{p\times\ell}$ the submatrix made with the columns from $V_\gamma$ associated to the corresponding cosines. In the same way, denoted by $W_\ell \in \mathbb{R}^{m\times\ell}$ the corresponding subset of columns in $W_\gamma$, and by $C_\ell \in \mathbb{R}^{\ell\times\ell}$ the reduction of $C_\gamma$ to the selected cosines. With these selected principal angles and vectors, we suggest considering the following reduced Schur complement preconditioner

$$S_\ell^{-1} = \alpha(B^TB)^{-1/2}\left(I_m - W_\ell\left(\frac{1}{\alpha}C_\ell^{-1}V_\ell^T\Lambda_\gamma V_\ell C_\ell^{-1} + I_\ell\right)^{-1}W_\ell^T\right)(B^TB)^{-1/2}. \tag{6.86}$$

In the next section, we introduce and illustrate a spectral preconditioner which uses reduced spectral information.

### 6.4.1 The spectral preconditioner with reduced spectral information

Among the two alternatives presented in Chapter 4 for preconditioning, we consider the cheapest one, built from (4.4) viz.

$$\mathcal{P}_\ell^{-1} = \begin{bmatrix} A_\ell^{-1} & 0 \\ 0 & S_\ell^{-1} \end{bmatrix},$$

where

$$A_\ell^{-1} = Y_\ell (Y_\ell^T A Y_\ell)^{-1} Y_\ell^T + \frac{1}{\alpha} I_n$$

and

$$S_\ell^{-1} = \alpha (B^T B)^{-1/2} \left( I_m - K_\ell \left( \frac{1}{\alpha} Y_\ell^T A Y_\ell + K_\ell^T K_\ell \right)^{-1} K_\ell^T \right) (B^T B)^{-1/2}, \tag{6.87}$$

which is equivalent to (6.86) in the case where $p \leq m$ and all cosines are nonzero. As the matrix $Y_\ell \in \mathbb{R}^{p \times \ell}$ does not necessarily contain the orthonormal set of the $p$ eigenvectors associated to the eigenvalues in A below $\gamma$, the rank-$\ell$ update in (6.87) is written as

$$\frac{1}{\alpha} Y_\ell^T A Y_\ell + K_\ell^T K_\ell.$$

We now illustrate the benefits of this last proposition on the previous test example. With a choice of $\gamma = \lambda_{\max}(A)/100 \approx 3.8\,10^{-2}$ and $\alpha = 1.16$ (see Section 3.1), the dimension of the invariant subspace $\mathcal{I}m(U_\gamma)$ is $p = 42$. The number of cosines of the principal angles (between $\mathcal{I}m(U_\gamma)$ and $\mathcal{I}m(B)$) less than $2\sqrt{\gamma/\alpha}$ is reduced to $\ell = 22$. Let us define $Y_\ell = U_\gamma V_\ell$, that represents the selected principal vectors in $\mathcal{I}m(U_\gamma)$, and $K_\ell = (B^T B)^{-1/2} B^T Y_\ell$.
Figure 6.4 shows the convergence profile of MINRES preconditioned with $\mathcal{P}_\ell$, built from the 22 selected smallest cosines. We can see that linear convergence is well established, as before, despite the reduction by about half of the size of the low rank update in the expression of $\mathcal{P}_1$. The scaled residual in $\mathcal{P}_\ell^{-1}$-norm is reduced to $10^{-8}$ after 88 iterations, while the reduction of the scaled residual in $\mathcal{P}_1^{-1}$-norm to $10^{-8}$ that was obtained after 58 iterations (as shown in Figure 4.2, for $\gamma = \lambda_{\max}(A)/100$). For sake of comparison, we also show in Figure 6.4 the convergence profile of MINRES preconditioned with $\mathcal{P}_1$ built up with the 22 smallest eigenvalues in $A$. We can see that the information carried out by the 22 selected principal angles is stronger, with respect to preconditioning issues, than the information carried out by an invariant subspace of the same size associated to the 22 smallest eigenvalues.

As we have mentioned before, (6.87) is equivalent to (6.86) only in the case where $p \leq m$ and all cosines are nonzero, and we would like to comment on

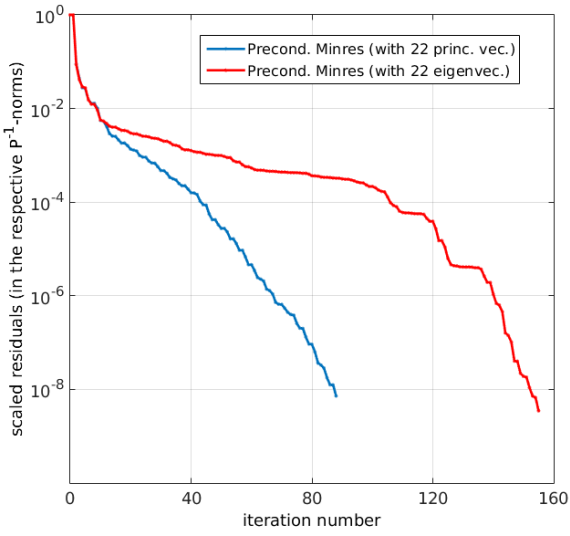



Figure 6.4 – $\mathcal{P}_\ell$ versus $\mathcal{P}_1$. The convergence profiles for the 22 smallest cosines and eigenvalues respectively.

the particular case where there exist rows in the rectangular matrix $\mathcal{C}$ that are zero. This can occur when $p > m$ and/or when there exists some orthogonality between $\mathcal{I}m(U_\gamma)$ and $\mathcal{I}m(B)$. In this case, equation (6.9) is not valid any more, since $\mathcal{J}\mathcal{J}^T$ incorporates zeros onto the diagonal and does not reduce to matrix $I_p$, and many different situations can occur for the interactions between the zeros or the ones in $\mathcal{J}\mathcal{J}^T$ and the $p \times p$ matrix $\mathcal{C}^\dagger V_\gamma^T \Lambda_\gamma V_\gamma \mathcal{C}^\dagger$. However, in the subcase where $p \leq m$ but some of the cosines are zero, looking at the more condensed formulation of $S_\ell^{-1}$ in (6.87), it is still possible to incorporate into matrix $Y_\ell$ those principal vectors associated to these particular zero cosines, and expect the preconditioner to raise the same properties. The only missing clue is to understand whether these vectors are of real importance or not.

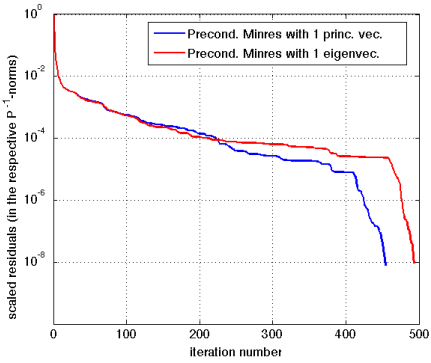
### 6.4.2 Reduced spectral information for various cut-off values

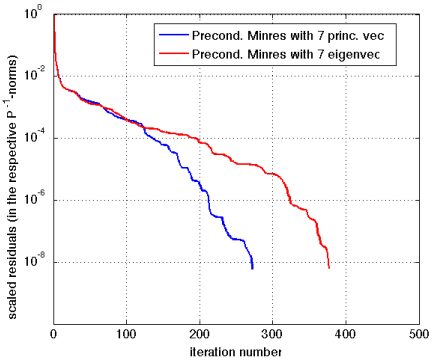
We remind that the cosines of the angles between $\mathcal{I}m(U_\gamma)$ and $\mathcal{I}m(B)$ do depend on the choice of the cut-off parameter $\gamma$ too, since different values for $\gamma$ change the dimension of $\mathcal{I}m(U_\gamma)$, and consequently the distribution of these cosines. Theorems 4.1 and 4.3 actually provide bounds on the macroscopic interactions between $A$ and $B$. Depending on the problem, there is a compromise to reach between the quantity of information embedded in $\mathcal{P}_1$ or $\mathcal{P}_2$, and the potential reduction in the rate of convergence of preconditioned Minres. The

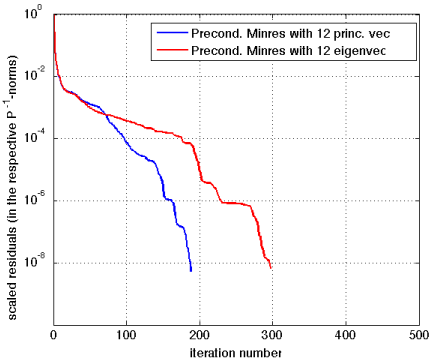
cosine selection more intimately targets the microscopic interactions within the given subspaces. This is a second feature, that might be considered on top of the first one, in order to still be able to reasonably target large values of $\gamma$ (to ensure nice asymptotic rates of convergence), while keeping the quantity of information to incorporate relatively small. Indeed, if one agrees with the fact that, in general, the distribution of these cosines is more or less well spread, then we might be left with few of them with regard to the size of the problem. The trade-off in the computational cost of all this, as well as the practical ways to extract the appropriate information, will surely depend on the application itself, and is devoted to further specific application oriented investigations. The main purpose in the discussion above is only to highlight the true components arising from the interactions between $A$ and $B$, and to investigate how best to incorporate these into a low rank update with valuable preconditioning effects.

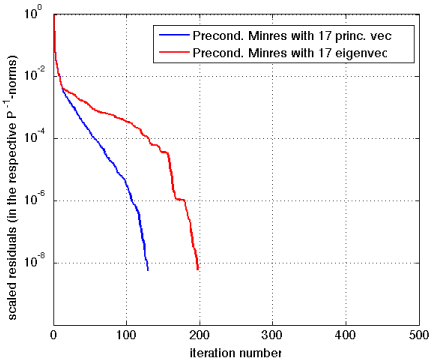
Now, in Figure 6.5, we analyse the behaviour of MINRES using the preconditioners $\mathcal{P}_1$ and $\mathcal{P}_\ell$ for various values of $\ell$ and $p$ to reach a scaled residual in $\mathcal{P}^{-1}$-norm below $10^{-8}$ with $\mathcal{P}$ being either $\mathcal{P}_\ell$ built from the $\ell$ smallest cosines, $\ell$ varying from 1 to 42, or $\mathcal{P}_1$ built from the $p$ smallest eigenvalues, $p$ also varying from 1 to 42. We first note that the number of iterations decreases with the value of $\ell$. Indeed, the preconditioner $\mathcal{P}_\ell$ or $\mathcal{P}_1$ is more efficient when we take into account a large quantity of information (the principal vectors for $\mathcal{P}_\ell$ or eigenvectors for $\mathcal{P}_1$, respectively). In Figure 6.5, the gap between the convergence curves of $\mathcal{P}_\ell$ and $\mathcal{P}_1$ decreases with the large values of $\ell$ and $p$ respectively, and finally the curves are similar for $\ell = p = 42$.

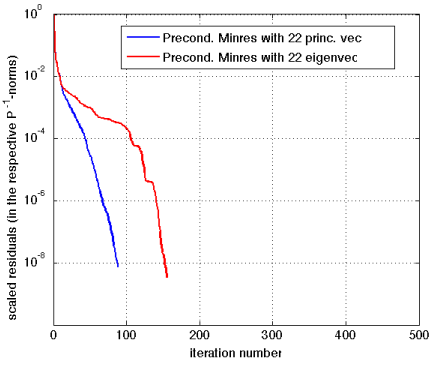
In Figure 6.6, we show the number of MINRES iterations needed to reach a scaled residual in $\mathcal{P}^{-1}$-norm below $10^{-8}$, $\mathcal{P}$ being either $\mathcal{P}_\ell$ built from the $\ell$ smallest cosines, $\ell$ varying from 1 to 42, or $\mathcal{P}_1$ built from the $p$ smallest eigenvalues, $p$ also varying from 1 to 42. We can see that the reduction in the number of iterations with increasing values of the cosines is faster than the one obtained with increasing eigenvalues. Additionally, we can see that this reduction roughly stabilizes after the first 25 smallest cosines, corresponding to the threshold value $\tau\sqrt{\gamma/\alpha}$ from the analysis above.

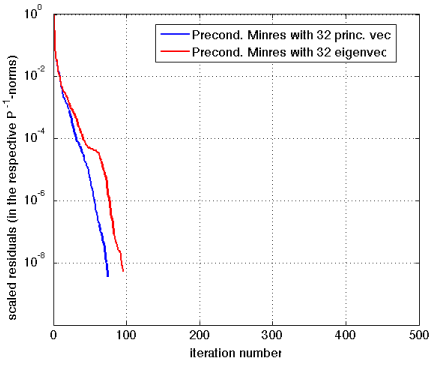

Precond. Minres with 1 princ. vec.
Precond. Minres with 1 eigenvec.
scaled residuals (in the respective P^-1-norms)
iteration number


Precond. Minres with 7 princ. vec
Precond. Minres with 7 eigenvec
scaled residuals (in the respective P^-1-norms)
iteration number


Precond. Minres with 12 princ. vec
Precond. Minres with 12 eigenvec
scaled residuals (in the respective P^-1-norms)
iteration number


Precond. Minres with 17 princ. vec
Precond. Minres with 17 eigenvec
scaled residuals (in the respective P^-1-norms)
iteration number


Precond. Minres with 22 princ. vec
Precond. Minres with 22 eigenvec
scaled residuals (in the respective P^-1-norms)
iteration number


Precond. Minres with 32 princ. vec
Precond. Minres with 32 eigenvec
scaled residuals (in the respective P^-1-norms)
iteration number

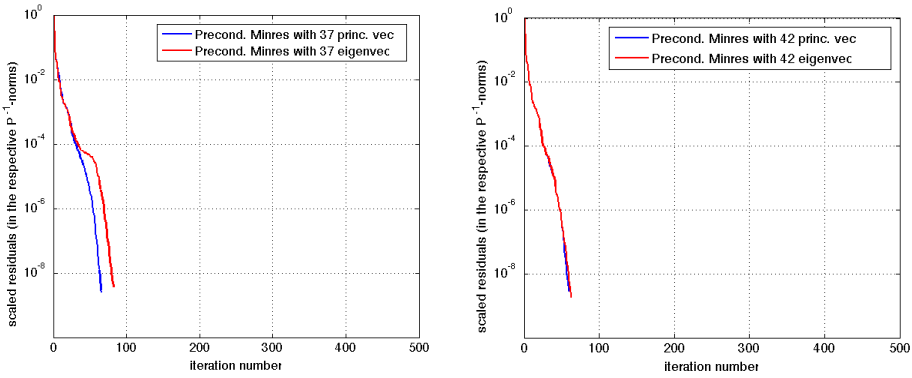



Figure 6.5 – $\mathcal{P}_\ell$ versus $\mathcal{P}_1$. Each subplot shows the convergence profiles for the $\ell$ smallest cosines and $p$ eigenvalues respectively. The value of $\ell$ and $p$ with $\ell = p$ is, from left to right: 1, 7, 12, 17, 22, 32, 37, 42.

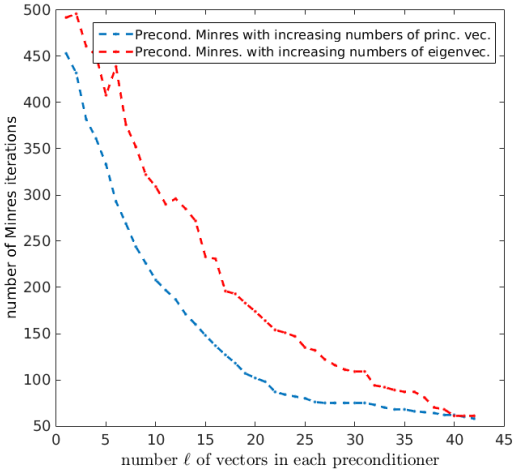



Figure 6.6 – $\mathcal{P}_\ell$ versus $\mathcal{P}_1$. Figure compares the number of MINRES iterations needed to reach $10^{-8}$ in both cases, with increasing numbers of cosines and eigenvalues respectively.

Chapter 7

# Practicalities

As we have seen in the previous chapters, spectral preconditioners are based on the approximation of the inverse of $(1,1)$ block $A$ and we have assumed that a first level of preconditioning $\mathcal{P}_0$ in (4.19), has been applied to the symmetric positive definite matrix $A$ so that the spectrum of $A$ is clustered, with relatively few very small eigenvalues. We have also initially assumed that we know these sets of small eigenvalues and associated eigenvectors of $A$. In this work, we come back to these two assumptions and we analyse some practical aspects on how to reach this assumptions. This chapter is divided into parts.

The first part of this chapter is based on how to combine a first level of preconditioning and the preconditioner developed in Giraud et al. (2006) and Golub et al. (2007), which combined with a Krylov method, enables to construct a Krylov basis of small dimension and very rich with respect to the eigeninformation linked to the smallest eigenvalues. This approach uses the general framework of Chebychev polynomials and Chebyshev filtering that we recall in Section 7.1.1. We then introduce the Chebychev polynomial preconditioner as in Giraud et al. (2006) and we illustrate in Section 7.1.2 the impact of Chebychev filtering on the spectrum of a matrix. In the next section, we derive our contribution on how one can condense these two preconditioners into a simple formulation, which can be used in practice.

In the second part, we focus on practical implementation of the approximation of the inverse of $A$ presented in Chapter 3 and used in Chapter 4 to introduce spectral preconditioners. Indeed, we consider the SLRU approach (3.4) where we set,

$$A_\gamma^{-1} = U_\gamma \Lambda_\gamma^{-1} U_\gamma^T + \frac{1}{\alpha} I_n \,. \tag{7.1}$$

with $\Lambda_\gamma \in \mathbb{R}^{p\times p}$ the diagonal matrix containing the $p$ eigenvalues less than $\gamma \in [\lambda_{\min}(A), \lambda_{\max}(A)]$. The columns of the rectangular matrix $U_\gamma \in \mathbb{R}^{n\times p}$ are the orthonormal sets of eigenvectors corresponding to $\Lambda_\gamma$. In Section 7.2.1,

we analyse from a theoretical point of view, the effect of these two levels of preconditioning (based on the condensed formulation developed in the first part of this chapter) on the spectrum of the preconditioned matrix. Finally, we see how one may derive in practice a good approximation of (7.1).

## 7.1 Extracting spectral information

This part recalls the general concepts of Chebyshev polynomials and introduce the Chebychev filtering, which will be used in Section 7.1.2 to extract prior spectral information from the matrix $A$ eventually preconditioned with a first level of preconditioning so as to cluster better the spectrum. The general framework of Chebyshev polynomials as proposed in the work Golub et al. (2007), is introduced with more details in the following sections.

### 7.1.1 General framework of Chebychev polynomial filtering

For any nonnegative integer $m$, we have defined the *Chebyshev polynomials* of degree $m$ in $w$ by the following two-term recurrence relation (Hageman and Young, 1981).

$$\begin{cases} T_0(w) = 1 \quad T_1(w) = w \\ T_{m+1}(w) = 2wT_m(w) - T_{m-1}(w) \quad m \geq 1, \end{cases} \tag{7.2}$$

or equivalently, we have that $T_m(w)$ may be expressed by

$$T_m(w) = \cos(m \arccos(w)) \quad \text{when} \quad w \in [-1, 1].$$

We have illustrated the first five Chebyshev polynomials in Figure 7.1 on the domain $[-1, 1]$. We consider a polynomial function of degree $m$ defined by

$$\mathcal{H}_m(w) = \frac{T_m(w)}{T_m(d)},$$

where $d > 1$ and we have the optimal properties of Chebyshev polynomials (see, e.g., Hageman and Young, 1981, Theorem 4.2.1),

$$\max_{w \in [-1,1]} |\mathcal{H}_m(w)| = \frac{1}{T_m(d)}.$$

For any polynomial $\mathcal{Q}_m(w)$ of degree $m$ or less such that $\mathcal{Q}_m(d) = 1$, if

$$\max_{w \in [-1,1]} |\mathcal{Q}_m(w)| \leq \max_{w \in [-1,1]} |\mathcal{H}_m(w)|,$$

then we have $\mathcal{Q}_m(w) = \mathcal{H}_m(w)$. We have that $\mathcal{H}_m(w)$ is the unique solution of

$$\min_{\mathcal{Q}_m} \max_{w \in [-1,1]} |\mathcal{Q}_m(w)|.$$

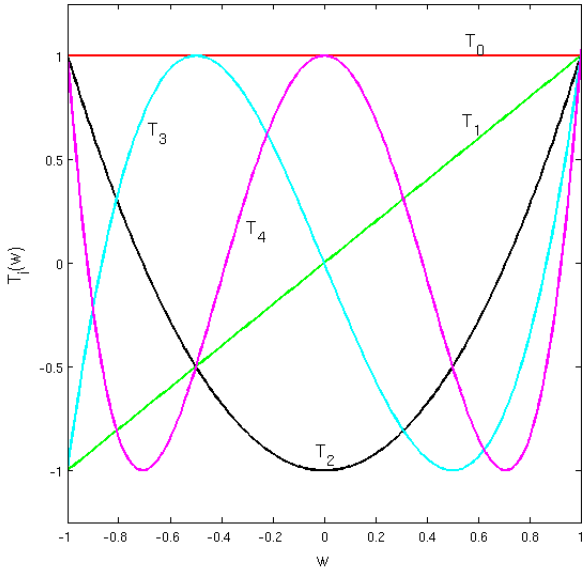



Figure 7.1 – We illustrate the first five Chebyshev polynomials with the x-axis between $-1$ and 1.

by Theorem 4.8 given by Saad (2011). We now define a linear transformation, which transforms the interval $[\gamma, \lambda_{\max}(A)]$ into $[-1, 1]$ :

$$\begin{array}{rccc} w_\gamma & : & \mathbb{R} & \to & \mathbb{R} \\ & & \lambda & \mapsto & w_\gamma(\lambda) = (\lambda_{\max}(A) + \gamma - 2\lambda)/(\lambda_{\max}(A) - \gamma) \end{array} \tag{7.3}$$

and we note $d_\gamma = w_\gamma(0) = \frac{\lambda_{\max}(A)+\gamma}{\lambda_{\max}(A)-\gamma} > 1$. In Golub et al. (2007), the authors introduced the filtering polynomial $\mathcal{F}_m(\lambda)$ defined by

$$\mathcal{F}_m(\lambda) = \frac{T_m(w_\gamma(\lambda))}{T_m(d_\gamma)}, \tag{7.4}$$

which has minimum upper bound value on the interval $[\gamma, \lambda_{\max}(A)]$ by the previous optimal properties. Let $\gamma$, $\lambda_{\max}(A)$ and $\epsilon \ll 1$, we can fix the degree

$m$ of $T_m$ such that $1/|T_m(d_\gamma)| < \epsilon$, which implies that

$$\begin{aligned}\max_{\lambda\in[\gamma,\lambda_{\max}]} |\mathcal{F}_m(\lambda)| &= \max_{\lambda\in[\gamma,\lambda_{\max}]} \left|\frac{T_m(w_\gamma(\lambda))}{T_m(d_\gamma)}\right| \\ &= \frac{1}{|T_m(d_\gamma)|} \max_{\lambda\in[\gamma,\lambda_{\max}]} \left|T_m(w_\gamma(\lambda))\right| \\ &< \epsilon \max_{\lambda\in[\gamma,\lambda_{\max}]} \left|T_m(w_\gamma(\lambda))\right|.\end{aligned}$$

By a property of Chebyshev polynomials (see, e.g., Saad, 2011, p.109), the maximum of the Chebyshev polynomial $T_m$ in $[-1, 1]$ is 1 and we then have

$$\max_{\lambda\in[\gamma,\lambda_{\max}]} |\mathcal{F}_m(\lambda)| < \epsilon. \tag{7.5}$$

We now consider a vector $w \in \mathbb{R}^n$ and we analyse the effect of the filtering polynomial on this vector. To see that, we construct a filtered vector as

$$w_f = \mathcal{F}_m(A)w = U_\gamma \mathcal{F}_m(\Lambda_\gamma)U_\gamma^T w + \tilde{U}_\gamma \mathcal{F}_m(\tilde{\Lambda}_\gamma)\tilde{U}_\gamma^T w, \tag{7.6}$$

where $\tilde{\Lambda}_\gamma \in \mathbb{R}^{(n-p)\times(n-p)}$ the diagonal matrix containing the $n-p$ eigenvalues greater than $\gamma \in [\lambda_{\min}(A), \lambda_{\max}(A)]$, and the columns of the rectangular matrix $\tilde{U}_\gamma \in \mathbb{R}^{n\times(n-p)}$ are the orthonormal sets of eigenvectors corresponding to $\tilde{\Lambda}_\gamma$. We pre-multiply the filtered vector (7.6) by $\tilde{U}_\gamma^T$ and we take the 2-norm. Using the Cauchy-Schwartz inequality, we obtain

$$\|\tilde{U}_\gamma^T w_f\|_2 \leq \|\mathcal{F}_m(\tilde{\Lambda}_\gamma)\|_2 \|\tilde{U}_\gamma^T w\|_2.$$

and by the vector norm properties (see, e.g., Golub and Van Loan, 1996, Section 2.2.2) and (7.5), we have

$$\begin{aligned}\|\tilde{U}_\gamma^T w_f\|_2 &\leq \sqrt{n-p} \quad \|\mathcal{F}_m(\tilde{\Lambda}_\gamma)\|_\infty \|\tilde{U}_\gamma^T w\|_2 \\ &< \sqrt{n-p} \quad \epsilon \quad \|\tilde{U}_\gamma^T w\|_2.\end{aligned}$$

with $\|\mathcal{F}_m(\tilde{\Lambda}_\gamma)\|_\infty = \max_{\lambda\in[\gamma,\lambda_{\max}]} |\mathcal{F}_m(\lambda)|$. This equation explains that the components of $w_f$ with respect to the invariant subspace $\tilde{U}_\gamma$ are reduced of factor $\sqrt{n-p} \quad \epsilon$.

In Figure 7.2, we illustrate the different behaviours of Chebyshev filtering $\mathcal{F}_{16}$ on $[\lambda_{\min}(A), \gamma]$ and $[\gamma, \lambda_{\max}(A)]$ respectively. Fixing the degree at $m = 16$, we obtain $\epsilon$ equal to $10^{-4}$ on $[\gamma, \lambda_{\max}(A)]$. The left figure shows that in the interval $[\lambda_{\min}(A), \gamma]$, the Chebyshev filtering is constant to 1 and decreases rapidly to zero when $\lambda$ is close to $\gamma$. Whereas in the interval $[\gamma, \lambda_{\max}(A)]$, the Chebychev filtering in right figure is equi-oscillating around zero and the relation (7.5) implies that the amplitude of oscillations is equal to $\epsilon$.

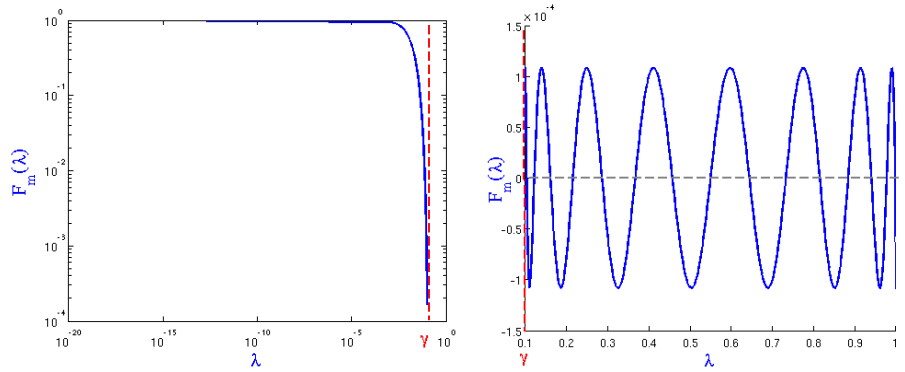



Figure 7.2 – On left, the Chebyshev filtering $\mathcal{F}_{16}$ on $[\lambda_{min}(A), \gamma]$ and at right, the Chebyshev filtering $\mathcal{F}_{16}$ on $[\gamma, \lambda_{\max}(A)]$.

### 7.1.2 Chebyshev polynomial preconditioner

The application of the *Chebyshev filter* as a preconditioner consists in solving approximatively a symmetric and positive definite linear system $Ax = b$ with the preconditioner defined by

$$\mathcal{P}_{\mathcal{F}}^{-1} := Q_{m-1}(A) = A^{-1}(I_n - \mathcal{F}_m(A)), \tag{7.7}$$

where $Q_{m-1}(A)$ is a matrix polynomial of degree less than or equal to $m-1$ (see, e.g., Hageman and Young, 1981, p.7). This preconditioner helps the CG method to generate a low dimensional Krylov basis that is very rich with respect to the smallest eigenvalues and associated eigenvectors. This spectral information can be used to build spectral approximations of the inverse of $A$ and of the Schur complement introduced in Section 3.1 and Section 3.2 respectively. This approach is developed in details in Golub et al. (2007) and we only introduce a simple example to show the interest.

We consider a symmetric positive definite matrix of order $n = 300$ randomly generated by the `Matlab` function `sprandsym`. The spectrum of the matrix has 10 eigenvalues less than $\frac{\lambda_{\max}(A)}{100}$. The Lanczos algorithm (see Algorithm 2, Section 2.1) combined with the preconditionner $\mathcal{P}_{\mathcal{F}}^{-1}$ in (7.7), plays the role of spectral filter and gives a basis of Krylov subspace $V$, which is a good approximation of eigenvectors of $A$ associated to the few small eigenvalues of $A$. As indicated in Figure 7.3 (left-hand plot), the values computed by $V^T AV$ plotting in blue are a good approximation of eigenvalues of $A$ less than $\gamma$. The right-hand plot represents the $\frac{\|Av_i - \lambda_i v_i\|}{|\lambda_i|}$ for each vector $v_i$ in $V$ and shows the accuracy of each vector.

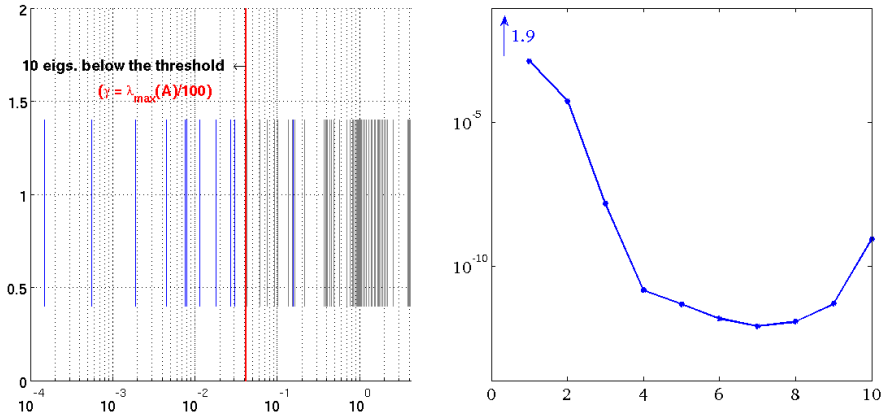



Figure 7.3 – Chebychev-based Krylov method.

### 7.1.3 Combining a first level preconditioner with Chebyshev polynomial preconditioner

Now, we explicitly develop our theoretical contribution on the application of the preconditioner $\mathcal{P}_{\mathcal{F}}^{-1}$ in (7.7), combined with a first level of preconditioning on the linear system denoted now

$$A_0 x = b,$$

where $A_0$ is symmetric and positive definite. As in Section 4.5, we consider a symmetric positive definite matrix $M$ given in a factorized form, $M = R^T R$ with $R \in \mathbb{R}^{n \times n}$. Let us first rewrite the preconditioned system

$$M^{-1} A_0 x = M^{-1} b$$

in a symmetrized manner as

$$A_1 y = c, \tag{7.8}$$

where $A_1 = R^{-T} A_0 \; R^{-1}$, $y = R \; x$ and $c = R^{-T} b$. We draw attention on the fact that we use a different notation for the initial system with first level preconditioner, compared to Section 4.5. The goal of which is to simplify and improve the clarity of this Section.

In the following theorem, we derive a general formulation that combines the two levels of preconditioning into one.

**Theorem 7.1** Let the linear system $A_0\ x = b$ and a symmetric positive definite preconditioner $M$ for $A_0$, given in a factorized form $R^T R$ with $R \in \mathbb{R}^{n\times n}$. Let the Chebyshev filter preconditioner $\mathcal{P}_{\mathcal{F}_1}^{-1}$ defined in (7.7) as $Q_{m-1}(A_1)$ with $A_1 = R^{-T} A_0\ R^{-1}$ and the preconditioned system $A_1 y = c$ with $y = R\ x$ and $c = R^{-T} b$. Then the combination of the two levels of preconditioning

$$\mathcal{P}_{\mathcal{F}_1}^{-1} A_1 y = \mathcal{P}_{\mathcal{F}_1}^{-1} c \tag{7.9}$$

is equivalent to

$$\mathcal{P}_{\mathcal{F}_0}^{-1} A_0 x = \mathcal{P}_{\mathcal{F}_0}^{-1} b,$$

where $\mathcal{P}_{\mathcal{F}_0}^{-1} = Q_{m-1}(M^{-1}A_0)M^{-1}$.

Note that the subscript $i$ in the preconditioner $\mathcal{P}_{\mathcal{F}_i}$ makes reference to the linear system with the matrix $A_i$.

**Proof.** The preconditioner $\mathcal{P}_{\mathcal{F}_1}^{-1}$ associated to the system $A_1 y = c$ is a matrix polynomial $Q_{m-1}(A_1)$ of the degree $(m-1)$ defined as

$$Q_{m-1}(A_1) = \alpha_0 I_n + \alpha_1 A_1 + \ldots + \alpha_{m-1} A_1^{m-1},$$

where $\{\alpha_i\}_{i=0}^{m-1}$ is a set of numbers (see, e.g., Hageman and Young, 1981, p. 7). We note that,

$$A_1^k = (R^{-T} A_0 R^{-1})(R^{-T} A_0 R^{-1}) \ldots (R^{-T} A_0 R^{-1}), \tag{7.10}$$

and premultiplying (7.10) by $RR^{-1}$, we get

$$\begin{aligned} A_1^k &= RR^{-1}(R^{-T} A_0 R^{-1})(R^{-T} A_0 R^{-1}) \ldots (R^{-T} A_0 R^{-1}) \\ &= R(M^{-1}A_0)^k R^{-1}, \end{aligned}$$

which implies that the preconditioner $\mathcal{P}_{\mathcal{F}_1}^{-1}$ can be expressed as

$$RQ_{m-1}(M^{-1}A_0)R^{-1}.$$

We replace this expression in (7.9) and by simplifying, we obtain

$$\begin{aligned} \mathcal{P}_{\mathcal{F}_1}^{-1} A_1 y = \mathcal{P}_{\mathcal{F}_1}^{-1} c \quad &\Leftrightarrow \quad RQ_{m-1}(M^{-1}A_0)R^{-1}A_1\ y = RQ_{m-1}(M^{-1}A_0)R^{-1}c \\ &\Leftrightarrow \quad Q_{m-1}(M^{-1}A_0)R^{-1}R^{-T}A_0R^{-1}Rx = Q_{m-1}(M^{-1}A_0)R^{-1}R^{-T}b \\ &\Leftrightarrow \quad Q_{m-1}(M^{-1}A_0)M^{-1}A_0x = Q_{m-1}(M^{-1}A_0)M^{-1}b. \end{aligned}$$

The two levels of preconditioning can be expressed by the formulation

$$\mathcal{P}_{\mathcal{F}_0}^{-1} = Q_{m-1}(M^{-1}A_0)M^{-1}.$$

□

Since we have a symmetric positive definite linear system, the conjugate gradient algorithm introduced in Section 2.1.2, is an algorithm of choice to solve it and can be used with the preconditioner in symmetric form

$$\mathcal{P}_{\mathcal{F}_0}^{-1} = R^{-1}Q_{m-1}(R^{-T}A_0R^{-1})R^{-T}. \tag{7.11}$$

This preconditioner is symmetric and positive definite as shown by the following result.

**Theorem 7.2** The preconditioner $R^{-1}Q_{m-1}(R^{-T}A_0R^{-1})R^{-T}$ is symmetric and positive definite.

**Proof.** We have that $Q_{m-1}(R^{-T}A_0R^{-1})$ is defined as a sum of symmetric matrices $I_n$, $R^{-T}A_0R^{-1}$, $(R^{-T}A_0R^{-1})^2$, …, $(R^{-T}A_0R^{-1})^{(m-1)}$, which implies that the preconditioner is symmetric. From relation (7.7), we obtain

$$\lambda Q_{m-1}(\lambda) = 1 - \mathcal{F}_m(\lambda), \tag{7.12}$$

for all eigenvalues $\lambda$ of $R^{-T}A_0R^{-1}$. By (7.12) and (7.5), we have that $\lambda Q_{m-1}(\lambda)$ belongs to $[1-\epsilon, 1+\epsilon]$ for all eigenvalues in $[\gamma, \lambda_{\max}(A_1)]$ (with $A_1 = R^{-T}A_0R^{-1}$). Since by construction, $\mathcal{F}_m(\lambda) \in [\epsilon, 1[$ for $\lambda \in ]0, \gamma]$, we get $\lambda Q(\lambda) \in ]0, 1-\epsilon]$ (since $\lambda Q(\lambda) = 1 - \mathcal{F}_m(\lambda)$) (see Figure 7.2). In short, for all $\lambda \in [0, \lambda_{\max}(A)]$, we have $Q_{m-1}(\lambda) > 0$ and we have that $Q_{m-1}(R^{-T}A_0R^{-1})$ is a positive definite matrix which implies by Meyer (2000), p.559, that for every nonzero $x \in \mathbb{R}^n$,

$$x^T Q_{m-1}(R^{-T}A_0R^{-1})x > 0.$$

Substituting $x$ by $R^{-T}z$, we obtain

$$\begin{aligned} x^T Q_{m-1}(R^{-T}A_0R^{-1})x > 0 \quad &\Leftrightarrow \quad (R^{-T}z)^T Q_{m-1}(R^{-T}A_0R^{-1})R^{-T}z > 0 \\ &\Leftrightarrow \quad z^T(R^{-1}Q_{m-1}(R^{-T}A_0R^{-1})R^{-T})z > 0 \end{aligned}$$

with nonzero $z \in \mathbb{R}^n$. We can deduce that the preconditioner is positive definite. □

## 7.2 Practical implementation of approximation of the inverse of the (1,1) block

As we have seen in Section 7.1.3, a first level preconditioner $M$ and the second level preconditioner, which is the Chebyshev filter preconditioner can be combined into one denoted by $\mathcal{P}_{\mathcal{F}_0}^{-1}$. In this section, we propose an approximation of the operator $A_\gamma^{-1}$ based on (7.1), which is associated to the preconditioned system

$$\mathcal{P}_{\mathcal{F}_0}^{-1} A_0 x = \mathcal{P}_{\mathcal{F}_0}^{-1} b. \tag{7.13}$$

with

$$\mathcal{P}_{\mathcal{F}_0}^{-1} = R^{-1} Q_{m-1}(R^{-T} A_0 R^{-1}) R^{-T} \tag{7.14}$$

and we denote by $A_0$ the initial matrix $A$. Note that the symmetric positive definite matrix $Q_{m-1}(R^{-T} A_0 R^{-1}) = Q_{m-1}(A_1)$ with the notation $A_1 := R^{-T} A_0 R^{-1}$, can be factorized in a square root form,

$$Q_{m-1}(A_1) = S_r^2, \tag{7.15}$$

with a symmetric matrix $S_r \in \mathbb{R}^{n\times n}$ implying in (7.14), the following factorization

$$\begin{aligned} \mathcal{P}_{\mathcal{F}_0}^{-1} &= R^{-1} S_r^2 R^{-T} \\ &= N^{-1} N^{-T} \end{aligned}$$

where $N = S_r^{-1} R \in \mathbb{R}^{n\times n}$. Denoting the preconditioned matrix by

$$A_2 := N^{-T} A_0 N^{-1},$$

one obtains the following relation between $A_2$ and $A_1$,

$$\begin{aligned} A_2 &= S_r R^{-T} A_0 R^{-1} S_r \\ &= S_r A_1 S_r. \end{aligned} \tag{7.16}$$

By (7.15), the matrix $S_r$ has the same eigenvectors as those of $A_1$. Since $S_r$ is symmetric, the matrix $S_r$ is diagonalizable in the same orthonormal basis of eigenvectors as this of $A_1$. We then have that $S_r$ and $A_1$ are said to commute (see Lancaster and Tismenetsky, 1985, Proposition 2) and we obtain by (7.15)

$$\begin{aligned} A_2 &= S_r^2 A_1 \\ &= Q_{m-1}(A_1) A_1. \end{aligned} \tag{7.17}$$

In Section 7.2.1, based on relation (7.17), we analyse the link between the eigenvalues of $A_1$ and the eigenvalues of $A_2$, which will be used in Section 7.2.2, to propose and analyse an good approximation of the operator $A_\gamma^{-1}$.

### 7.2.1 Eigenvalue distribution of matrix with two levels of preconditioning

For clarification, the notations that we will use for the following analysis are given in Table 7.1 (we recall that $N = S_r^{-1}R$). The first column holds the notation of the diagonal matrices with various levels of preconditioning. The notations used for the eigenvectors and the eigenvalues for each matrix are defined in the second and third columns, respectively. The columns of matrices in the second column are the orthogonal set of eigenvectors corresponding to the diagonal matrices in third column containing the eigenvalues.

| matrix | eigenvectors | eigenvalues |
|---|---|---|
| $A_0$ | $U$ | $\Lambda = \text{diag}\{\lambda_i\}_{i=1}^n$ |
| $A_1 = R^{-T}A_0R^{-1}$ | $V$ | $\Theta = \text{diag}\{\theta_i\}_{i=1}^n$ |
| $A_2 = N^{-T}A_0N^{-1} = S_r^2A_1$ | $W$ | $\Delta = \text{diag}\{\delta_i\}_{i=1}^n$ |

Table 7.1 – Table of notations

With (7.16), the next theorem gives the relation between the eigenvalues of the two preconditioned matrices $A_1$ and $A_2$.

**Theorem 7.3** Let $\Theta \in \mathbb{R}^{n\times n}$ the diagonal matrix containing the $n$ eigenvalues of $A_1$ and the columns of the matrix $V \in \mathbb{R}^{n\times n}$ form the orthogonal set of the eigenvectors corresponding to $\Theta$. Then the diagonal matrix containing the $n$ eigenvalues of $A_2$ is defined as

$$\Delta = Q_{m-1}(\Theta)\Theta,$$

and the associated eigenvectors denoted by the columns of $W$ can be given by the eigenvectors of $A_1$.

**Proof.** By (7.17), the eigendecomposition of $A_1$ and the orthogonality of $V$, we have that

$$\begin{aligned} A_2 &= Q_{m-1}(A_1)A_1 \\ &= Q_{m-1}(V\Theta V^T)V\Theta V^T \\ &= V\left(Q_{m-1}(\Theta)\Theta\right)V^T. \end{aligned}$$

□

Note that all the eigenvectors of $A_1$ are necessary eigenvectors of $A_2$, while the opposite is not always true. Indeed, we consider two eigenvalues $\theta_i$ and $\theta_j$

of $A_1$ and $v_i$ and $v_j$ the corresponding eigenvectors respectively. We consider the case where $Q_{m-1}(\theta_i)\theta_i = Q_{m-1}(\theta_j)\theta_j$ and for instance, $\frac{1}{\sqrt{2}}(v_i + v_j)$ is an eigenvector of $A_2$ but not for $A_1$. Indeed, we have

$$\begin{aligned} A_1\left(\frac{1}{\sqrt{2}}(v_i+v_j)\right) &= \frac{1}{\sqrt{2}}\left(A_1v_i + A_1v_j\right) \\ &= \frac{1}{\sqrt{2}}\left(\theta_i v_i + \theta_j v_j\right) \\ &\neq \theta\frac{1}{\sqrt{2}}\left(v_i+v_j\right), \end{aligned}$$

where $\theta \in \mathbb{R}$.

### 7.2.2 Approximation of the inverse of the (1,1) block

Let us consider the following expression

$$\tilde{A}_\gamma^{-1} = \frac{1}{\alpha}M^{-1} + P_k(P_k^TA_0P_k)^{-1}P_k^T, \tag{7.18}$$

where the columns of the matrix $P_k \in \mathbb{R}^{n\times k}$ form the set of $A_0$-conjugate directions computed in the conjugate gradient algorithm preconditioned (see Section 2.2) by $\mathcal{P}_{\mathcal{F}_0}^{-1} = N^{-1}N^{-T}$ to solve the linear system $A_0x = b$. By Proposition 6.7 in Saad (2011) when the Krylov subspace is invariant, we can express the basis which are the columns of $P_k$, as a linear combination of a subset of $k$ eigenvectors $V_k \in \mathbb{R}^{n\times k}$ (since by Theorem 7.3, the eigenvectors of $A_1$ are equal to the eigenvectors of $A_2$) of the symmetrically preconditioned system $A_2$ such that

$$P_k = N^{-1}V_k\beta_k, \tag{7.19}$$

where $\beta_k \in \mathbb{R}^{k\times k}$ is an invertible matrix. Similarly to the eigendecomposition of the matrix $A$ in (3.3), we split the spectrum of $A_1$ in two parts, with $\Theta_\gamma \in \mathbb{R}^{p\times p}$ the diagonal matrix containing the $p$ eigenvalues less than a given positive number $\gamma \in [\theta_{\min}(A_1), \theta_{\max}(A_1)]$, and with $\tilde{\Theta}_\gamma \in \mathbb{R}^{(n-p)\times(n-p)}$ the diagonal matrix containing all the other $(n-p)$ eigenvalues. The columns of $V_\gamma \in \mathbb{R}^{n\times p}$ and $\tilde{V}_\gamma \in \mathbb{R}^{n\times(n-p)}$ are the orthonormal sets of eigenvectors corresponding to $\Theta_\gamma$ and $\tilde{\Theta}_\gamma$ respectively. We have the following spectral relations

$$A_1V_\gamma = V_\gamma\Theta_\gamma \quad \text{and} \quad A_1\tilde{V}_\gamma = \tilde{V}_\gamma\tilde{\Theta}_\gamma, \tag{7.20}$$

and by Theorem 7.3, we deduce

$$A_2V_\gamma = V_\gamma\Delta_\gamma \quad \text{and} \quad A_2\tilde{V}_\gamma = \tilde{V}_\gamma\tilde{\Delta}_\gamma \tag{7.21}$$

with $\Delta_\gamma = Q_{m-1}(\Theta_\gamma)\Theta_\gamma$ and $\tilde{\Delta}_\gamma = Q_{m-1}(\tilde{\Theta}_\gamma)\tilde{\Theta}_\gamma$.

Assuming that the initial residual is not orthogonal to the eigenvectors of $A_2$ corresponding to the eigenvalues $\Delta_\gamma$ and considering that happy breakdown has occured within the Lanczos process, we know that the Krylov subspace is an invariant subspace and that the set of eigenvectors $V_k$ in (7.19) incorporates necessarily all of $V_\gamma$, so that

$$V_k = \left[V_\gamma, \tilde{V}_\gamma \gamma_p\right]$$

with $\gamma_p \in \mathbb{R}^{(n-p)\times(k-p)}$. By (7.19) we have

$$P_k = N^{-1}\left[V_\gamma, \tilde{V}_\gamma \gamma_p\right]\beta_k. \tag{7.22}$$

Substituting (7.22) for $P_k$ in (7.18) with $P = P_k$, we can see that

$$\begin{aligned} P_k^T A_0 P_k &= \beta_k^T \begin{bmatrix} V_\gamma^T \\ \gamma_p^T \tilde{V}_\gamma^T \end{bmatrix} N^{-T} A_0 N^{-1} \left[V_\gamma \quad \tilde{V}_\gamma \gamma_p\right] \beta_k \\ &= \beta_k^T \begin{bmatrix} V_\gamma^T \\ \gamma_p^T \tilde{V}_\gamma^T \end{bmatrix} A_2 \left[V_\gamma \quad \tilde{V}_\gamma \gamma_p\right] \beta_k, \end{aligned}$$

since $A_2 = N^{-T} A_0 N^{-1}$. We have that $V_\gamma^T A_2 \tilde{V}_\gamma = V_\gamma^T \tilde{V}_\gamma \tilde{\Delta}_\gamma = 0$ by (7.21), implying that

$$P_k^T A_0 P_k = \beta_k^T \begin{bmatrix} V_\gamma^T A_2 V_\gamma & 0 \\ 0 & \gamma_p^T \tilde{V}_\gamma^T A_2 \tilde{V}_\gamma \gamma_p \end{bmatrix} \beta_k. \tag{7.23}$$

Observe that the matrix $P_k^T A_0 P_k$ is block diagonal such that the inverse of $P_k^T A_0 P_k$ is also block diagonal including the inverse of each block. We get the rewritting of (7.18) with two terms after elimination of the invertible $k \times k$ matrix $\beta_k$, such that

$$\begin{aligned} \tilde{A}_\gamma^{-1} &= \frac{1}{\alpha} M^{-1} + N^{-1} V_\gamma \left(V_\gamma^T A_2 V_\gamma\right)^{-1} V_\gamma^T N^{-T} \\ &\quad + N^{-1} \tilde{V}_\gamma \gamma_p \left(\gamma_p^T \tilde{V}_\gamma^T A_2 \tilde{V}_\gamma \gamma_p\right)^{-1} \gamma_p^T \tilde{V}_\gamma^T N^{-T}. \end{aligned}$$

Using spectral relations (7.21), we obtain

$$\begin{aligned} \tilde{A}_\gamma^{-1} &= \frac{1}{\alpha} M^{-1} + N^{-1} V_\gamma \Delta_\gamma^{-1} V_\gamma^T N^{-T} \\ &\quad + N^{-1} \tilde{V}_\gamma \gamma_p \left(\gamma_p^T \tilde{\Delta}_\gamma \gamma_p\right)^{-1} \gamma_p^T \tilde{V}_\gamma^T N^{-T}. \end{aligned} \tag{7.24}$$

The following result establishes a first expression of the preconditioned matrix $\tilde{A}_\gamma^{-1}A_0$ and explicitly shows the effect of the preconditioner $\tilde{A}_\gamma^{-1}$ on $A_0$.

**Theorem 7.4** Let the matrix $A_0 \in \mathbb{R}^{n\times n}$ and the preconditioner $\tilde{A}_\gamma^{-1}$ defined by (7.24). Then we obtain, with the notations introduced above,

$$\tilde{A}_\gamma^{-1}A_0 = R^{-1}\left(\frac{1}{\alpha}A_1 + V_\gamma\Theta_\gamma^{-1}V_\gamma^T A_1 + S_r\tilde{V}_\gamma\gamma_p\left(\gamma_p^T\tilde{\Delta}_\gamma\gamma_p\right)^{-1}\gamma_p^T\tilde{V}_\gamma^T S_r^T A_1\right)R.$$

**Proof.** By $N = S_r^{-1}R$ and $M = R^TR$, we first write (7.24) as

$$\begin{aligned}\tilde{A}_\gamma^{-1} = \frac{1}{\alpha}R^{-1}R^{-T} &+ R^{-1}S_rV_\gamma\Delta_\gamma^{-1}V_\gamma^TS_r^TR^{-T}\\ &+ R^{-1}S_r\tilde{V}_\gamma\gamma_p\left(\gamma_p^T\tilde{\Delta}_\gamma\gamma_p\right)^{-1}\gamma_p^T\tilde{V}_\gamma^TS_r^TR^{-T}.\end{aligned}$$

If we apply the operator $\tilde{A}_\gamma^{-1}$ on $A_0$, we have the following formulation

$$\begin{aligned}\tilde{A}_\gamma^{-1}A_0 = \frac{1}{\alpha}R^{-1}R^{-T}A_0 &+ R^{-1}S_rV_\gamma\Delta_\gamma^{-1}V_\gamma^TS_r^TR^{-T}A_0\\ &+ R^{-1}S_r\tilde{V}_\gamma\gamma_p\left(\gamma_p^T\tilde{\Delta}_\gamma\gamma_p\right)^{-1}\gamma_p^T\tilde{V}_\gamma^TS_r^TR^{-T}A_0\end{aligned}$$

or, equivalently,

$$\begin{aligned}\tilde{A}_\gamma^{-1}A_0 = R^{-1}\Bigg(&\frac{1}{\alpha}R^{-T}AR^{-1} + S_rV_\gamma\Delta_\gamma^{-1}V_\gamma^TS_r^TR^{-T}AR^{-1}\\ &+ S_r\tilde{V}_\gamma\gamma_p\left(\gamma_p^T\tilde{\Delta}_\gamma\gamma_p\right)^{-1}\gamma_p^T\tilde{V}_\gamma^TS_r^TR^{-T}AR^{-1}\Bigg)R.\end{aligned}$$

Using $A_1 = R^{-T}A_0R^{-1}$, we obtain

$$\begin{aligned}\tilde{A}_\gamma^{-1}A_0 = &\\ R^{-1}\left(\frac{1}{\alpha}A_1 + S_rV_\gamma\Delta_\gamma^{-1}V_\gamma^TS_r^TA_1 + S_r\tilde{V}_\gamma\gamma_p\left(\gamma_p^T\tilde{\Delta}_\gamma\gamma_p\right)^{-1}\gamma_p^T\tilde{V}_\gamma^TS_r^TA_1\right)R.\end{aligned} \tag{7.25}$$

Considering the eigendecomposition of $A_1$, i.e., $A_1 = V_\gamma\Theta_\gamma V_\gamma^T + \tilde{V}_\gamma\tilde{\Theta}_\gamma\tilde{V}_\gamma^T$ and by (7.15) with $Q_{m-1}(A_1)$ symmetric positive definite, we obtain

$$\begin{aligned} S_r V_\gamma &= (Q_{m-1}(A_1))^{1/2} V_\gamma \\ &= \left(Q_{m-1}\left(V_\gamma \Theta_\gamma V_\gamma^T + \tilde{V}_\gamma \tilde{\Theta}_\gamma \tilde{V}_\gamma^T\right)\right)^{1/2} V_\gamma \\ &= V_\gamma (Q_{m-1}(\Theta_\gamma))^{1/2}, \end{aligned}$$

and by Theorem 7.3,

$$S_r V_\gamma = V_\gamma \Theta_\gamma^{-1/2} \Delta_\gamma^{1/2}. \tag{7.26}$$

Replacing (7.26) in the second therm in (7.25), we obtain

$$\begin{aligned} S_r V_\gamma \Delta_\gamma^{-1} V_\gamma^T S_r^T A_1 &= V_\gamma \Theta_\gamma^{-1/2} \Delta_\gamma^{1/2} \Delta_\gamma^{-1} \Delta_\gamma^{1/2} \Theta_\gamma^{-1/2} V_\gamma^T A_1 \\ &= V_\gamma \Theta_\gamma^{-1/2} \Theta_\gamma^{-1/2} V_\gamma^T A_1. \end{aligned}$$

and we conclude the proof. □

By the similarity transformation $R$, this theorem tells us that the eigenvalues of $\tilde{A}_\gamma^{-1} A_0$ are similar to the eigenvalues of

$$\frac{1}{\alpha} A_1 + V_\gamma \Theta_\gamma^{-1} V_\gamma^T A_1 + S_r \tilde{V}_\gamma \gamma_p \left(\gamma_p^T \tilde{\Delta}_\gamma \gamma_p\right)^{-1} \gamma_p^T \tilde{V}_\gamma^T S_r^T A_1, \tag{7.27}$$

where the first two terms in (7.27) represent the effect of the spectral preconditioner

$$\frac{1}{\alpha} I_n + V_\gamma \Theta_\gamma^{-1} V_\gamma^T$$

on $A_1$. This preconditioner is known as the SLRU approach (see Giraud et al., 2006) associated with spectral information of $A_1$ defined as in (7.1). The following theorem rewrites the previous expression of $\tilde{A}_\gamma^{-1} A_0$, so as to ease the last part of the analysis.

**Theorem 7.5** Let the matrix $A_0 \in \mathbb{R}^{n \times n}$ and the preconditioner $\tilde{A}_\gamma^{-1}$ defined by (7.24). Then we obtain

$$\tilde{A}_\gamma^{-1} A_0 = R^{-1} \left(\frac{1}{\alpha} A_1 + V_\gamma V_\gamma^T + \left(\tilde{V}_\gamma \tilde{\Theta}_\gamma^{-1/2}\right) Y \left(\tilde{\Theta}_\gamma^{1/2} \tilde{V}_\gamma^T\right)\right) R,$$

where $Y = \left(\tilde{\Delta}_\gamma^{1/2} \gamma_p \left(\gamma_p^T \tilde{\Delta}_\gamma \gamma_p\right)^{-1} \gamma_p^T \tilde{\Delta}_\gamma^{1/2}\right)$.

**Proof.** By (7.20), the first part $\frac{1}{\alpha}A_1 + V_\gamma \Theta_\gamma^{-1} V_\gamma^T A_1$ in (7.27) is equal to $\frac{1}{\alpha}A_1 + V_\gamma \Theta_\gamma^{-1}\Theta_\gamma V_\gamma^T = \frac{1}{\alpha}A_1 + V_\gamma V_\gamma^T$. Similarly to equation (7.26), we have $S_r \tilde{V}_\gamma = \tilde{V}_\gamma \tilde{\Theta}_\gamma^{-1/2}\tilde{\Delta}_\gamma^{1/2}$ such that we obtain the formulation

$$A_\gamma^{-1}A_0 = R^{-1}\left(\frac{1}{\alpha}A_1 + V_\gamma V_\gamma^T + \tilde{V}_\gamma\tilde{\Theta}_\gamma^{-1/2}\tilde{\Delta}_\gamma^{1/2}\gamma_p\left(\gamma_p^T\tilde{\Delta}_\gamma\gamma_p\right)^{-1}\gamma_p^T\tilde{\Delta}_\gamma^{1/2}\tilde{\Theta}_\gamma^{-1/2}\tilde{V}_\gamma^T A_1\right)R,$$

or equivalently, using (7.20),

$$A_\gamma^{-1}A_0 = R^{-1}\left(\frac{1}{\alpha}A_1 + V_\gamma V_\gamma^T + \tilde{V}_\gamma\tilde{\Theta}_\gamma^{-1/2}\left(\tilde{\Delta}_\gamma^{1/2}\gamma_p\left(\gamma_p^T\tilde{\Delta}_\gamma\gamma_p\right)^{-1}\gamma_p^T\tilde{\Delta}_\gamma^{1/2}\right)\tilde{\Theta}_\gamma^{1/2}\tilde{V}_\gamma^T\right)R.$$

□

The second formulation of $\tilde{A}_\gamma^{-1}A_0$ shows explicitly the effect of the spectral preconditioner applied to $A_0$. By the first and the second terms of $\tilde{A}_\gamma^{-1}A_0$, the smallest eigenvalues of $A_1$ belonging to $[0,\gamma]$ are transformed into $(\frac{1}{\alpha}\theta_i + 1)$, which meets the expectation as we have seen in Section 3.1. In the meantime, the largest eigenvalues of $A_1$ belonging to $[\gamma, \theta_{\max}(A_1)]$ are approximately equal to $\frac{1}{\alpha}\theta_i$ plus a bounded corrective third term involved by a similarity transformation the eigenvalues of

$$Y = \tilde{\Delta}_\gamma^{1/2}\gamma_p\left(\gamma_p^T\tilde{\Delta}_\gamma\gamma_p\right)^{-1}\gamma_p^T\tilde{\Delta}_\gamma^{1/2}.$$

The matrix $Y$ is positive semidefinite since $Y$ corresponds to the orthogonal projection (i.e. $Y = Y^2$ and $Y = Y^T$) implying that the eigenvalues of $Y$ are equal to 0 or 1. By these considerations, we have that the eigenvalues of $A_\gamma^{-1}A_0$ are isolated away from zero. Finally, this third term is active only on the invariant subspace linked to the largest eigenvalues of $A_1$, (in the interval $[\gamma, \theta_{\max}(A_1)]$), because of the products with $\tilde{V}_\gamma$ and $\tilde{V}_\gamma^T$ on both sides. The internal part is similar to the orthogonal projector $Y$, the similitude being given by the square root of $\tilde{\Theta}_\gamma$, and therefore this third term may only shift marginally the largest eigenvalues towards $+\infty$, and at maxima by a factor of $\sqrt{(\gamma/\theta_{\max}(A_1))}$.

Now, in practice, we may expect that before "happy breakdown" is actually reached, the Krylov subspace $P_k$ obtained in the chebyshev-preconditioned CG will already be rich enough to incorporate all of those eigenvectors in $V_\gamma$, and that therefore the resulting preconditionner that we shall get will have spectral properties close to what we have analysed above. This is what we shall illustrate now on a pratical example. We consider the previous test example (see Section 3.1) with $\gamma = \lambda_{\max}(A_0)/100$. The eigenvalues of $A_\gamma^{-1}A_0$ and $\tilde{A}_\gamma^{-1}A_0$ belong to

$$\left[4.287297471632787\ 10^{-2}, 3.281045885810809\right]$$

and

$$\left[4.2872974717\mathbf{48417}\ 10^{-2}, 3.2810458858\mathbf{15332}\right]$$

respectively and the relative error between eigenvalues of $A_\gamma^{-1}A_0$ and $\tilde{A}_\gamma^{-1}A_0$ is equal to $8.14\ 10^{-11}$. In Figure 7.4, we illustrate the effect of preconditioner $\tilde{A}_\gamma^{-1}$ on the eigenvalues of $A_0$. The top figure illustrates the eigenvalue distribution of $A_0$ such that the eigenvalues $\lambda_i$ less than $\gamma$ are represented with a green star and by Theorem 7.5 are transformed to $\frac{\lambda_i}{\alpha} + 1$ with $\alpha = 1.16$ represented respectively at the bottom figure.

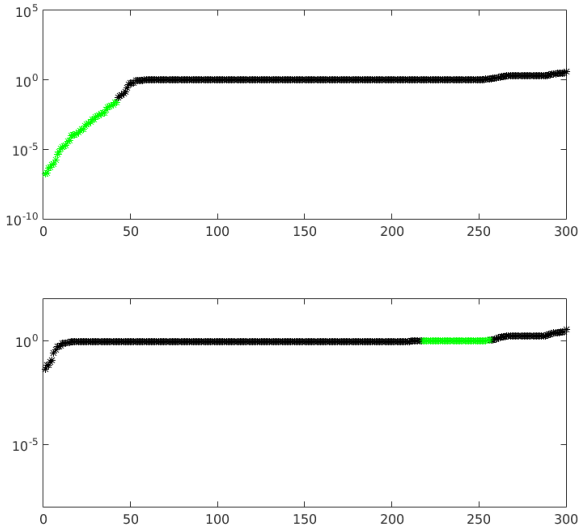



Figure 7.4 – On top, the eigenvalues distribution of $A_\gamma^{-1}A_0$ and at the bottom, the eigenvalues distribution of $\tilde{A}_\gamma^{-1}A_0$.

Figure 7.5 compares the number of MINRES iteration needed to reach $10^{-8}$ in both cases, for $\mathcal{P}_1$ defined with exact spectral information and $\tilde{\mathcal{P}}_1$ defined as

$$\tilde{\mathcal{P}}_1 = \begin{bmatrix} \tilde{A}_\gamma & 0 \\ 0 & \tilde{S}_\gamma \end{bmatrix}$$

where $\tilde{S}_\gamma = B^T\tilde{A}_\gamma^{-1}B$ with $\tilde{A}_\gamma^{-1}$ as given in (7.18). We can see that the number of iterations with the preconditioner $\tilde{\mathcal{P}}_1$ is equal to 68, while that in the case where the preconditioner is $\mathcal{P}_1$, reaches 69.

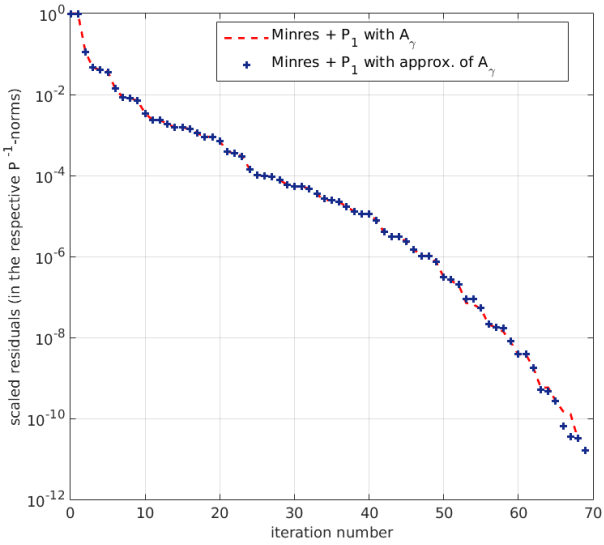



Figure 7.5 – Convergence profiles of preconditioned Minres with preconditioners $\mathcal{P}_1$ and $\tilde{\mathcal{P}}_1$.

# Conclusions and Perspectives

This work focusses on block diagonal preconditioning for KKT systems and SQD systems to ensure fast convergence of Krylov iterative solvers such as MINRES. We have introduced two preconditioners that approximate the "ideal" preconditioner proposed by Murphy et al. (2000) for KKT systems and by Gould and Simoncini (2009) for SQD systems, that are able to reduce the spectrum of the preconditioned saddle-point matrix.

The two alternatives of block diagonal preconditioners that we propose are based on good approximations of the invariant subspace associated with the ill-conditioned part of the $(1,1)$ block $A$, which can then be exploited to construct efficient approximations for the Schur complement. Our first purpose is to analyse the benefits of such an approach for preconditioning Krylov solvers, and the theoretical analysis given in the various theorems actually shows that convergence of preconditioned MINRES can be accelerated.

The two specific block diagonal preconditioners actually incorporate a low rank update of the Schur complement itself, that can be superimposed on top of a first level preconditioning that reduces as much as possible the dimension of the invariant subspace containing the ill-conditioned part of the resulting $(1,1)$ block. We have shown that, when a first level of preconditioning is used with our preconditioners for the KKT systems, an effective new version of our preconditioners can be applied on the initial system.

The theoretical results and practical considerations contained in this work show that the proposed technique is a good complement to a first level of preconditioning whenever this is not sufficient to obtain fast convergence for MINRES. The efficiency, as well as the practical feasibility of our preconditioners, will for sure depend on the application itself. The ideal conditions would be to have the benefits of a first level of preconditioning that manages to reduce the ill-conditioning within the constraint equations and tightly clusters the spectrum in the $(1,1)$ block as well.

An important part of our analysis gives some insights on the interaction between the $(1,1)$ block $A$ and the constraints $(1,2)$ block, showing in which circumstances the bad conditioning contained in $A$ effectively spoils the convergence of MINRES. We have also refined the bounds given by Rusten and Winther (1992) on the eigenvalues of a KKT matrix and investigated how best

to incorporate the appropriate spectral information with valuable preconditioning effects on saddle-point systems.

One of our perspectives is to complete the spectral approach on $A$ with a similar approach to address the bad conditioning of $B$. Moreover, we could also analyse the cost and the amortization of our technique for the solution of several saddle-point systems involving the same matrix and multiple right-hand sides. Another perspective is to study and compare SQD systems arising in interior-point methods. Preliminary results show that the preconditioners with efficient implementation could be competitive compared to constraint preconditioners. Another interesting point that is still open for future work is the extension of our study to more general cases with less restrictive assumptions on the blocks of the saddle-point matrices.

# List of Tables

# List of Figures

# Appendices

Appendix A

# Tools of linear algebra

## A.1 The singular value decomposition

*The singular value decomposition* (SVD) of the matrix $A$ is given by the next theorem.

**Theorem A.1** If $A \in \mathbb{R}^{m\times n}$, then there exist orthogonal matrices

$$U = [u_1, \ldots, u_m] \in \mathbb{R}^{m\times m} \text{ and } V = [v_1, \ldots, v_n] \in \mathbb{R}^{n\times n}$$

such that

$$U^T AV = \Sigma = \text{diag}(\sigma_1, \ldots, \sigma_p) \in \mathbb{R}^{m\times n}, \quad p = \min\{m, n\},$$

where $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0$

**Proof.** See, e.g., (Golub and Van Loan, 2013, p.76) □

The singular values of the matrix $A$ and the left and right singular vectors of $A$ are defined by the following result.

**Corollary A.2** If $U^T AV = \Sigma$ is the SVD of $A \in \mathbb{R}^{m\times n}$ $(m \geq n)$, then

$$Av_i = \sigma_i u_i \qquad \text{and} \qquad A^T u_i = \sigma_i v_i \quad \text{for} \quad i = 1, \ldots, n.$$

**Proof.** See, e.g., (Golub and Van Loan, 2013, p.77) □

We can deduce from this corollary that

$$A^T A v_i = \sigma_i^2 v_i \quad \text{and} \quad A^T A u_i = \sigma_i^2 u_i \quad \text{for} \quad i = 1, \ldots, n.$$

Indeed, the singular values of $A$ are the positive square roots of the eigenvalues of $AA^T$ or $A^T A$.

**Corollary A.3** If $A \in \mathbb{R}^{m \times n}$, then $\|A\|_2 = \sigma_1$.

**Proof.** See, e.g., (Golub and Van Loan, 2013, p.77) □

## A.2 The principal angles and the associated principal vectors

Let $\mathcal{F}$ and $\mathcal{G}$ be subspaces in $\mathbb{R}^n$ whose dimensions satisfy

$$p = \dim(\mathcal{F}) \geq \dim(\mathcal{G}) = q \geq 1.$$

The *principal angles* $\{\theta_k\}_{k=1}^q$ between these two subspaces and the associated *principal vectors* $\{f_k\}_{k=1}^q$ and $\{g_k\}_{k=1}^q$ are defined recursively by

$$\cos(\theta_k) = f_k^T g_k = \max_{\substack{f \in \mathcal{F}, \|f\|_2 = 1 \\ f^T[f_1, \ldots, f_{k-1}] = 0}} \max_{\substack{g \in \mathcal{G}, \|g\|_2 = 1 \\ g^T[g_1, \ldots, g_{k-1}] = 0}} f^T g,$$

where $0 \leq \theta_1 \leq \ldots \leq \theta_q \leq \frac{\pi}{2}$.

A singular value decomposition for computing cosines of principal angles between the subspaces $\mathcal{F}$ and $\mathcal{G}$ can be formulated as follows. Let the columns of matrices $Q_{\mathcal{F}} \in \mathbb{R}^{n \times p}$ and $Q_{\mathcal{G}} \in \mathbb{R}^{n \times q}$ form orthonormal bases for the subspaces $\mathcal{F}$ and $\mathcal{G}$, respectively. The thin SVD of $Q_{\mathcal{F}}^T Q_{\mathcal{G}}$ is

$$U^T Q_{\mathcal{F}}^T Q_{\mathcal{G}} V = \operatorname{diag}(\sigma_1, \ldots, \sigma_q),$$

where $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_q \geq 0$, $U \in \mathbb{R}^{p \times q}$ and $V \in \mathbb{R}^{q \times q}$ both have orthonormal columns. Then principal angles can be computed as

$$\sigma_k = \cos(\theta_k), \quad k = 1, \ldots, q,$$

where $0 \leq \theta_1 \leq \ldots \leq \theta_q \leq \frac{\pi}{2}$, while the vectors $\{f_k\}_{k=1}^q$ and $\{g_k\}_{k=1}^q$ are the associated principal vectors computed by

$$f_k = (Q_{\mathcal{F}} U)(:, k), \quad g_k = (Q_{\mathcal{G}} V)(:, k), \quad k = 1, \ldots, q.$$

## A.3 The Sherman-Morrison-Woodbury formula

The following formula is defined as *the Sherman-Morrison formula* and gives a convenient expression for the inverse of the matrix $(A+UV^T)$ where $A \in \mathbb{R}^{n\times n}$ invertible, $U$ and $V \in \mathbb{R}^{n\times k}$,

$$(A+UV^T)^{-1} = A^{-1} - A^{-1}U(I_n + V^TA^{-1}U)^{-1}V^TA^{-1}.$$

This expression shows how to compute the inverse of a rank-k correction $UV^T$ to a matrix $A$ in a rank-k correction of the inverse. Note that $A+UV^T$ is invertible if and only if $I_n + V^TA^{-1}U$ is invertible.

## A.4 The cosine decomposition

**Theorem A.4** Let $P$ a unitary matrix of dimension $m+p=k+q$, then there exist unitary matrices $U$, $V$, $W$, $Z$ of dimensions $m$, $p$, $k$, $q$ respectively, so that

$$\begin{pmatrix} U^\star & 0 \\ 0 & V^\star \end{pmatrix} P \begin{pmatrix} W & 0 \\ 0 & Z \end{pmatrix} =$$

$$\left(\begin{array}{ccc|ccc} I & & & 0_s^\star & & \\ & C & & & S & \\ & & 0_c & & & I \\ \hline 0_s & & & I & & \\ & S & & & -C & \\ & & I & & & 0_c^\star \end{array}\right) \begin{array}{c} r \\ s \\ m-r-s \\ p-k+r \\ s \\ k-r-s \end{array}$$

$$\begin{array}{ccccccc} r & s & k-r-s & p-k+r & s & m-r-s \end{array}$$

$$C = \operatorname{diag}(\alpha_{r+1}, \ldots, \alpha_{r+s}), \qquad 1 > \alpha_{r+1} \geqslant \ldots \geqslant \alpha_{r+s} > 0,$$
$$S = \operatorname{diag}(\beta_{r+1}, \ldots, \beta_{r+s}), \qquad 0 < \beta_{r+1} \leqslant \ldots \leqslant \beta_{r+s} < 1,$$

$$C^2 + S^2 = I.$$

**Proof.** See, e.g., (Paige and Saunders, 1981, p.403) □

Appendix B

# Proof of Theorems

## B.1 Proof of Theorem 3.1 of Chapter 3

First note that the matrices $S$ and $S_\gamma$ are symmetric and positive definite, hence nonsingular, by definition of $A$ and $A_\gamma$ respectively, and by the full column rank property of $B \in \mathbb{R}^{n\times m}$ (see, e.g., Golub and Van Loan, 2013, Section 4.2.1). The eigenvalue problem $S_\gamma^{-1}Sx = \lambda x$ is then equivalent to the generalized eigenvalue problem:

$$Sx = \lambda S_\gamma x, \tag{B.1}$$

that is, $\lambda(S_\gamma^{-1}S) = \lambda(S, S_\gamma) = \{\nu_i\}_{i=1}^m$.

The first part of the proof transforms problem (B.1) into two successively equivalent generalized eigenvalue problems. We define the matrices

$$S^{(a)} = (B^TB)^{-1/2}S(B^TB)^{-1/2} = Q^TA^{-1}Q \tag{B.2}$$

and

$$S_\gamma^{(a)} = (B^TB)^{-1/2}S_\gamma(B^TB)^{-1/2} = Q^TA_\gamma^{-1}Q, \tag{B.3}$$

where $Q = B(B^TB)^{-1/2} \in \mathbb{R}^{n\times m}$. The first equality of each equation and the nonsingularity of $(B^TB)^{-1/2}$ guarantee that $\lambda(S, S_\gamma) = \lambda(S^{(a)}, S_\gamma^{(a)}) = \{\nu_i\}_{i=1}^m$. To exploit the second equality of each equation, consider the matrix $K \in \mathbb{R}^{m\times n}$ defined as

$$K = Q^TU = [Q^TU_\gamma,\ Q^T\tilde{U}_\gamma] = [K_\gamma,\ \tilde{K}_\gamma], \tag{B.4}$$

where $Q$ satisfies $Q^TQ = I_m$ by definition, $U = [U_\gamma,\ \tilde{U}_\gamma]$ is the orthogonal matrix of the eigendecomposition (3.3), $K_\gamma$ is the operator used in (3.13) and we set $\tilde{K}_\gamma = Q^T\tilde{U}_\gamma$. The columns of $K^T$ are orthonormal, implying that

$K_\gamma K_\gamma^T + \tilde{K}_\gamma \tilde{K}_\gamma^T = I_m$. If we now complete the matrix $K^T$ by $m-n$ orthonormal columns to provide an orthogonal matrix of $\mathbb{R}^{n\times n}$, and if we apply the CS decomposition as in Appendix A or Paige and Saunders (1981), Section 4, one can guarantee the existence of orthogonal matrices $V_\gamma \in \mathbb{R}^{p\times p}$, $\tilde{V}_\gamma \in \mathbb{R}^{(n-p)\times(n-p)}$ and $W \in \mathbb{R}^{m\times m}$ such that

$$V_\gamma^T K_\gamma^T W = \mathcal{C} = \mathrm{diag}(c_1, \ldots, c_r) \in \mathbb{R}^{p\times m}, \qquad r = \min\{p, m\}, \tag{B.5}$$

and

$$\tilde{V}_\gamma^T \tilde{K}_\gamma^T W = \mathcal{S} = \mathrm{diag}(s_1, \ldots, s_q) \in \mathbb{R}^{(n-p)\times m}, \qquad q = \min\{n-p, m\}, \tag{B.6}$$

where $\mathcal{C}^T\mathcal{C} + \mathcal{S}^T\mathcal{S} = I_m$. The singular values $c_i$ and $s_i$ of $K_\gamma^T$ and $\tilde{K}_\gamma^T$ respectively, are cosines and sines satisfying (without loss of generality)

$$1 \geq c_1 \geq \ldots \geq c_r \geq 0 \qquad \text{and} \qquad 0 \leq s_1 \leq \ldots \leq s_q \leq 1. \tag{B.7}$$

In principle, the $\min\{r, q\}$ of these values correspond to the cosines and sines of the principal angles between $\mathcal{I}m(B)$ and $\mathcal{I}m(U_\gamma)$, the other values being equal to either 0 or 1, depending on the dimensions $p, m$ and $n$. Extracting $K_\gamma$ and $\tilde{K}_\gamma$ from (B.5) and (B.6) yields

$$K_\gamma = W\mathcal{C}^T V_\gamma^T \qquad \text{and} \qquad \tilde{K}_\gamma = W\mathcal{S}^T \tilde{V}_\gamma^T. \tag{B.8}$$

We now come back to the second equalities of equations (B.2) and (B.3) and use the expressions (3.3) and (3.4) to derive expressions for $A^{-1}$ and $A_\gamma^{-1}$ respectively. One obtains, by (B.4) and (B.8),

$$\begin{aligned} S^{(a)} &= KU^T(U_\gamma \Lambda_\gamma^{-1} U_\gamma^T + \tilde{U}_\gamma \tilde{\Lambda}_\gamma^{-1} \tilde{U}_\gamma^T) U K^T \\ &= (K_\gamma U_\gamma^T + \tilde{K}_\gamma \tilde{U}_\gamma^T)(U_\gamma \Lambda_\gamma^{-1} U_\gamma^T + \tilde{U}_\gamma \tilde{\Lambda}_\gamma^{-1} \tilde{U}_\gamma^T)(U_\gamma K_\gamma^T + \tilde{U}_\gamma \tilde{K}_\gamma^T) \\ &= K_\gamma \Lambda_\gamma^{-1} K_\gamma^T + \tilde{K}_\gamma \tilde{\Lambda}_\gamma^{-1} \tilde{K}_\gamma^T \\ &= W S^{(b)} W^T, \end{aligned} \tag{B.9}$$

where

$$S^{(b)} = \mathcal{C}^T V_\gamma^T \Lambda_\gamma^{-1} V_\gamma \mathcal{C} + \mathcal{S}^T \tilde{V}_\gamma^T \tilde{\Lambda}_\gamma^{-1} \tilde{V}_\gamma \mathcal{S}, \tag{B.10}$$

and, similarly,

$$\begin{aligned} S_\gamma^{(a)} &= KU^T(U_\gamma \Lambda_\gamma^{-1} U_\gamma^T + \frac{1}{\alpha} I_n) U K^T \\ &= (K_\gamma U_\gamma^T + \tilde{K}_\gamma \tilde{U}_\gamma^T)(U_\gamma \Lambda_\gamma^{-1} U_\gamma^T + \frac{1}{\alpha} I_n)(U_\gamma K_\gamma^T + \tilde{U}_\gamma \tilde{K}_\gamma^T) \\ &= K_\gamma(\Lambda_\gamma^{-1} + \frac{1}{\alpha} I_p) K_\gamma^T + \frac{1}{\alpha} \tilde{K}_\gamma \tilde{K}_\gamma^T \\ &= W S_\gamma^{(b)} W^T, \end{aligned} \tag{B.11}$$

where

$$S_\gamma^{(b)} = \mathcal{C}^T V_\gamma^T \left(\Lambda_\gamma^{-1} + \frac{1}{\alpha} I_p\right) V_\gamma \mathcal{C} + \frac{1}{\alpha} \mathcal{S}^T \mathcal{S}. \tag{B.12}$$

The last equalities of (B.9) and (B.11) together with the nonsingularity of $W$ then guarantee that $\lambda(S^{(a)}, S_\gamma^{(a)}) = \lambda(S^{(b)}, S_\gamma^{(b)}) = \{\nu_i\}_{i=1}^m$.

For the second part of the proof, consider, for a non-zero vector $x \in \mathbb{R}^m$, the generalized Rayleigh quotient

$$\nu(x) = \frac{x^T S^{(b)} x}{x^T S_\gamma^{(b)} x}. \tag{B.13}$$

We show that $\nu(x)$ belongs to the interval defined in (3.15), hence implying the desired result. First observe that

$$\lambda(\Lambda_\gamma^{-1}) \in \left[\frac{1}{\gamma}, \frac{1}{\lambda_{\min}(A)}\right] \qquad \text{and} \qquad \lambda(\tilde{\Lambda}_\gamma^{-1}) \in \left[\frac{1}{\lambda_{\max}(A)}, \frac{1}{\gamma}\right], \tag{B.14}$$

by definition of $\Lambda_\gamma$ and $\tilde{\Lambda}_\gamma$. By (B.10) and (B.12), we can write

$$\nu(x) = \frac{x^T \mathcal{C}^T V_\gamma^T \Lambda_\gamma^{-1} V_\gamma \mathcal{C} x + x^T \mathcal{S}^T \tilde{V}_\gamma^T \tilde{\Lambda}_\gamma^{-1} \tilde{V}_\gamma \mathcal{S} x}{x^T \mathcal{C}^T V_\gamma^T \left(\Lambda_\gamma^{-1} + \frac{1}{\alpha} I_p\right) V_\gamma \mathcal{C} x + \frac{1}{\alpha} x^T \mathcal{S}^T \mathcal{S} x},$$

yielding the following equalities

$$\begin{aligned} \nu(x) &= \frac{x^T \mathcal{C}^T V_\gamma^T (\alpha \Lambda_\gamma^{-1}) V_\gamma \mathcal{C} x}{x^T \mathcal{C}^T V_\gamma^T (\alpha \Lambda_\gamma^{-1} + I_p) V_\gamma \mathcal{C} x + x^T \mathcal{S}^T \mathcal{S} x} \\ &\quad + \frac{x^T \mathcal{S}^T \tilde{V}_\gamma^T (\alpha \tilde{\Lambda}_\gamma^{-1}) \tilde{V}_\gamma \mathcal{S} x}{x^T \mathcal{C}^T V_\gamma^T (\alpha \Lambda_\gamma^{-1} + I_p) V_\gamma \mathcal{C} x + x^T \mathcal{S}^T \mathcal{S} x} \end{aligned} \tag{B.15}$$

and

$$\begin{aligned} \frac{1}{\nu(x)} &= \frac{x^T \mathcal{C}^T V_\gamma^T (\alpha \Lambda_\gamma^{-1} + I_p) V_\gamma \mathcal{C} x}{x^T \mathcal{C}^T V_\gamma^T (\alpha \Lambda_\gamma^{-1}) V_\gamma \mathcal{C} x + x^T \mathcal{S}^T \tilde{V}_\gamma^T (\alpha \tilde{\Lambda}_\gamma^{-1}) \tilde{V}_\gamma \mathcal{S} x} \\ &\quad + \frac{x^T \mathcal{S}^T \mathcal{S} x}{x^T \mathcal{C}^T V_\gamma^T (\alpha \Lambda_\gamma^{-1}) V_\gamma \mathcal{C} x + x^T \mathcal{S}^T \tilde{V}_\gamma^T (\alpha \tilde{\Lambda}_\gamma^{-1}) \tilde{V}_\gamma \mathcal{S} x}. \end{aligned} \tag{B.16}$$

Consider first the case where both $\mathcal{C}x$ and $\mathcal{S}x$ are non-zero vectors. Since $\alpha > 0$, the matrices $V_\gamma^T (\alpha \Lambda_\gamma^{-1}) V_\gamma$ and $\tilde{V}_\gamma^T (\alpha \tilde{\Lambda}_\gamma^{-1}) \tilde{V}_\gamma$ are positive definite and each term of the numerator and the denominator in (B.15) and in (B.16) is positive. One can thus write

$$\begin{aligned} \nu(x) &\leq \frac{x^T \mathcal{C}^T V_\gamma^T (\alpha \Lambda_\gamma^{-1}) V_\gamma \mathcal{C} x}{x^T \mathcal{C}^T V_\gamma^T (\alpha \Lambda_\gamma^{-1} + I_p) V_\gamma \mathcal{C} x} + \frac{x^T \mathcal{S}^T \tilde{V}_\gamma^T (\alpha \tilde{\Lambda}_\gamma^{-1}) \tilde{V}_\gamma \mathcal{S} x}{x^T \mathcal{S}^T \mathcal{S} x} \\ &\leq 1 + \frac{\alpha}{\gamma}, \end{aligned} \tag{B.17}$$

using the second part of (B.14) and the fact that $x^T\mathcal{S}^T\mathcal{S}x = x^T\mathcal{S}^T\tilde{V}_\gamma^T\tilde{V}_\gamma\mathcal{S}x$. In the same way, we can write

$$\begin{aligned}\frac{1}{\nu(x)} &\leq \frac{x^T\mathcal{C}^TV_\gamma^T(\alpha\Lambda_\gamma^{-1})V_\gamma\mathcal{C}x}{x^T\mathcal{C}^TV_\gamma^T(\alpha\Lambda_\gamma^{-1})V_\gamma\mathcal{C}x} + \frac{x^T\mathcal{C}^TV_\gamma^TV_\gamma\mathcal{C}x}{x^T\mathcal{C}^TV_\gamma^T(\alpha\Lambda_\gamma^{-1})V_\gamma\mathcal{C}x} \\ &\quad + \frac{x^T\mathcal{S}^T\mathcal{S}x}{x^T\mathcal{S}^T\tilde{V}_\gamma^T(\alpha\tilde{\Lambda}_\gamma^{-1})\tilde{V}_\gamma\mathcal{S}x} \\ &\leq 1+\frac{\gamma}{\alpha}+\frac{\lambda_{\max}(A)}{\alpha}, \end{aligned} \tag{B.18}$$

where we use both parts of (B.14) and the fact that $x^T\mathcal{S}^T\mathcal{S}x = x^T\mathcal{S}^T\tilde{V}_\gamma^T\tilde{V}_\gamma\mathcal{S}x$. The bounds (B.17) and (B.18) imply the desired result for this first case. Now, from $\mathcal{C}^T\mathcal{C}+\mathcal{S}^T\mathcal{S}=I_m$, one cannot have both $\mathcal{C}x=0$ and $\mathcal{S}x=0$ at the same time for a non-zero $x\in\mathbb{R}^m$. Considering these particular cases separately, we can deduce from (B.14), (B.15) and (B.16) when $\mathcal{C}x=0$,

$$\begin{aligned}\nu(x) &= \frac{x^T\mathcal{S}^T\tilde{V}_\gamma^T(\alpha\tilde{\Lambda}_\gamma^{-1})\tilde{V}_\gamma\mathcal{S}x}{x^T\mathcal{S}^T\mathcal{S}x} \\ &\leq \frac{\alpha}{\gamma} \\ &\leq 1+\frac{\alpha}{\gamma}\end{aligned}$$

and

$$\begin{aligned}\frac{1}{\nu(x)} &= \frac{x^T\mathcal{S}^T\mathcal{S}x}{x^T\mathcal{S}^T\tilde{V}_\gamma^T(\alpha\tilde{\Lambda}_\gamma^{-1})\tilde{V}_\gamma\mathcal{S}x} \\ &\leq \frac{\lambda_{\max}(A)}{\alpha} \\ &\leq 1+\frac{\gamma}{\alpha}+\frac{\lambda_{\max}(A)}{\alpha}.\end{aligned}$$

Similarly, when $\mathcal{S}x=0$, we deduce from (B.15) and (B.16),

$$\begin{aligned}\nu(x) &= \frac{x^T\mathcal{C}^TV_\gamma^T(\alpha\Lambda_\gamma^{-1})V_\gamma\mathcal{C}x}{x^T\mathcal{C}^TV_\gamma^T(\alpha\Lambda_\gamma^{-1}+I_p)V_\gamma\mathcal{C}x} \\ &\leq 1 \\ &\leq 1+\frac{\alpha}{\gamma}\end{aligned}$$

and

$$\begin{aligned}\frac{1}{\nu(x)} &= \frac{x^T\mathcal{C}^T V_\gamma^T(\alpha\Lambda_\gamma^{-1}+I_p)V_\gamma\mathcal{C}x}{x^T\mathcal{C}^T V_\gamma^T(\alpha\Lambda_\gamma^{-1})V_\gamma\mathcal{C}x}\\ &\leq 1+\frac{\gamma}{\alpha}\\ &\leq 1+\frac{\gamma}{\alpha}+\frac{\lambda_{\max}(A)}{\alpha}.\end{aligned}$$

In both cases, the bounds in (B.17) and (B.18) are still valid, which ends the proof.

□

## B.2 Proof of Theorem 3.4 of Chapter 3

First note that the matrices $S$ and $S_\gamma$ are symmetric and positive definite, hence nonsingular, by the definition of $A$ and $A_\gamma$, respectively, and by the full column rank property of $B \in \mathbb{R}^{n\times m}$ (see, e.g., Golub and Van Loan, 2013, Section 4.2.1). The eigenvalue problem $S_\gamma^{-1}Sx = \lambda x$ is then equivalent to the generalized eigenvalue problem:

$$Sx = \lambda S_\gamma x, \tag{B.19}$$

that is, $\lambda(S_\gamma^{-1}S) = \lambda(S, S_\gamma) = \{\nu_i\}_{i=1}^m$.

The first part of the proof transforms problem (B.19) into two successively generalized eigenvalue problems. We define the matrices

$$S^{(a)} = (B^TB)^{-1/2}S(B^TB)^{-1/2} = Q^TA^{-1}Q + (B^TB)^{-1/2}C(B^TB)^{-1/2} \tag{B.20}$$

and

$$S_\gamma^{(a)} = (B^TB)^{-1/2}S_\gamma(B^TB)^{-1/2} = Q^TA_\gamma^{-1}Q + (B^TB)^{-1/2}C(B^TB)^{-1/2}, \tag{B.21}$$

where $Q = B(B^TB)^{-1/2} \in \mathbb{R}^{n\times m}$. Observe that the terms $Q^TA^{-1}Q$ and $Q^TA_\gamma^{-1}Q$ in (B.20) and (B.21) are similar to (B.2) and (B.3) respectively, implying that we can thus follow the similar steps in the proof of Theorem 3.1. Indeed, defining

$$K = Q^TU = [Q^TU_\gamma,\ Q^T\tilde{U}_\gamma] = [K_\gamma,\ \tilde{K}_\gamma], \tag{B.22}$$

where $Q$ satisfies $Q^TQ = I_m$ and using the CS Decomposition as in Appendix A or Paige and Saunders (1981), Section 4, we obtain

$$K_\gamma = W\mathcal{C}^TV_\gamma^T \qquad \text{and} \qquad \tilde{K}_\gamma = W\mathcal{S}^T\tilde{V}_\gamma^T. \tag{B.23}$$

Similarly to (B.9) and (B.11), we now come back to the second equalities of equations (B.20) and (B.21) and use the expressions (3.3) and (3.4) to derive expressions for $A^{-1}$ and $A_\gamma^{-1}$, respectively. One obtains, by (B.22) and (B.23),

$$S^{(a)} = WS^{(b)}W^T + (B^TB)^{-1/2}C(B^TB)^{-1/2}, \tag{B.24}$$

where

$$S^{(b)} = \mathcal{C}^T V_\gamma^T \Lambda_\gamma^{-1} V_\gamma \mathcal{C} + \mathcal{S}^T \tilde{V}_\gamma^T \tilde{\Lambda}_\gamma^{-1} \tilde{V}_\gamma \mathcal{S}, \tag{B.25}$$

and, similarly,

$$S_\gamma^{(a)} = WS_\gamma^{(b)}W^T + (B^TB)^{-1/2}C(B^TB)^{-1/2}, \tag{B.26}$$

where

$$S_\gamma^{(b)} = \mathcal{C}^T V_\gamma^T \left(\Lambda_\gamma^{-1} + \frac{1}{\alpha} I_p\right) V_\gamma \mathcal{C} + \frac{1}{\alpha}\mathcal{S}^T\mathcal{S}. \tag{B.27}$$

For the second part of the proof, consider, for a non-zero vector $x \in \mathbb{R}^m$, the generalized Rayleigh quotient

$$\nu(x) = \frac{x^T S^{(a)} x}{x^T S_\gamma^{(a)} x} \tag{B.28}$$

or, equivalently by (B.24) and (B.26),

$$\nu(x) = \frac{x^T W S^{(b)} W^T x + x^T (B^TB)^{-1/2} C (B^TB)^{-1/2} x}{x^T W S_\gamma^{(b)} W^T x + x^T (B^TB)^{-1/2} C (B^TB)^{-1/2} x}.$$

Setting $y = W^T x$ and $z = (B^TB)^{-1/2}x$ as non-zero vectors, we have

$$\nu(x) = \frac{y^T S^{(b)} y + z^T C z}{y^T S_\gamma^{(b)} y + z^T C z}. \tag{B.29}$$

We will show that $\nu(x)$ belongs to the interval defined in (3.26), hence implying the desired result. First observe that

$$\lambda(\Lambda_\gamma^{-1}) \in \left[\frac{1}{\gamma}, \frac{1}{\lambda_{\min}(A)}\right] \qquad \text{and} \qquad \lambda(\tilde{\Lambda}_\gamma^{-1}) \in \left[\frac{1}{\lambda_{\max}(A)}, \frac{1}{\gamma}\right], \tag{B.30}$$

by definition of $\Lambda_\gamma$ and $\tilde{\Lambda}_\gamma$. By (B.25) and (B.27), we can write

$$\nu(x) = \frac{y^T\mathcal{C}^T V_\gamma^T \Lambda_\gamma^{-1} V_\gamma \mathcal{C} y + y^T \mathcal{S}^T \tilde{V}_\gamma^T \tilde{\Lambda}_\gamma^{-1} \tilde{V}_\gamma \mathcal{S} y + z^T C z}{y^T\mathcal{C}^T V_\gamma^T \left(\Lambda_\gamma^{-1} + \frac{1}{\alpha} I_p\right) V_\gamma \mathcal{C} y + \frac{1}{\alpha} y^T \mathcal{S}^T \mathcal{S} y + z^T C z},$$

yielding the following equalities

$$\begin{aligned}
\nu(x) \quad = \quad & \frac{y^T\mathcal{C}^T V_\gamma^T (\alpha\Lambda_\gamma^{-1}) V_\gamma \mathcal{C} y}{y^T\mathcal{C}^T V_\gamma^T (\alpha\Lambda_\gamma^{-1} + I_p) V_\gamma \mathcal{C} y + y^T \mathcal{S}^T \mathcal{S} y + z^T \alpha C z} \\
& + \quad \frac{y^T \mathcal{S}^T \tilde{V}_\gamma^T (\alpha\tilde{\Lambda}_\gamma^{-1}) \tilde{V}_\gamma \mathcal{S} y}{y^T\mathcal{C}^T V_\gamma^T (\alpha\Lambda_\gamma^{-1} + I_p) V_\gamma \mathcal{C} y + y^T \mathcal{S}^T \mathcal{S} y + z^T \alpha C z} \\
& + \quad \frac{z^T \alpha C z}{y^T\mathcal{C}^T V_\gamma^T (\alpha\Lambda_\gamma^{-1} + I_p) V_\gamma \mathcal{C} y + y^T \mathcal{S}^T \mathcal{S} y + z^T \alpha C z}
\end{aligned} \tag{B.31}$$

and

$$\begin{aligned}\frac{1}{\nu(x)} &= \frac{y^T\mathcal{C}^T V_\gamma^T(\alpha\Lambda_\gamma^{-1}+I_p)V_\gamma\mathcal{C}y}{y^T\mathcal{C}^T V_\gamma^T(\alpha\Lambda_\gamma^{-1})V_\gamma\mathcal{C}y + y^T\mathcal{S}^T\tilde{V}_\gamma^T(\alpha\tilde{\Lambda}_\gamma^{-1})\tilde{V}_\gamma\mathcal{S}y + z^T\alpha Cz} \\ &+ \frac{y^T\mathcal{S}^T\mathcal{S}y}{y^T\mathcal{C}^T V_\gamma^T(\alpha\Lambda_\gamma^{-1})V_\gamma\mathcal{C}y + y^T\mathcal{S}^T\tilde{V}_\gamma^T(\alpha\tilde{\Lambda}_\gamma^{-1})\tilde{V}_\gamma\mathcal{S}y + z^T\alpha Cz} \\ &+ \frac{z^T\alpha Cz}{y^T\mathcal{C}^T V_\gamma^T(\alpha\Lambda_\gamma^{-1})V_\gamma\mathcal{C}y + y^T\mathcal{S}^T\tilde{V}_\gamma^T(\alpha\tilde{\Lambda}_\gamma^{-1})\tilde{V}_\gamma\mathcal{S}y + z^T\alpha Cz}. \quad \text{(B.32)}\end{aligned}$$

Consider first the case where both $\mathcal{C}y$ and $\mathcal{S}y$ are non-zero vectors. Since $\alpha > 0$, the matrices $V_\gamma^T(\alpha\Lambda_\gamma^{-1})V_\gamma$, $\tilde{V}_\gamma^T(\alpha\tilde{\Lambda}_\gamma^{-1})\tilde{V}_\gamma$ and $C$ are positive definite and each term of the numerator and the denominator in (B.31) and in (B.32) is positive. One can thus write

$$\begin{aligned}\nu(x) &\leq \frac{y^T\mathcal{C}^T V_\gamma^T(\alpha\Lambda_\gamma^{-1})V_\gamma\mathcal{C}y}{y^T\mathcal{C}^T V_\gamma^T(\alpha\Lambda_\gamma^{-1}+I_p)V_\gamma\mathcal{C}y} + \frac{y^T\mathcal{S}^T\tilde{V}_\gamma^T(\alpha\tilde{\Lambda}_\gamma^{-1})\tilde{V}_\gamma\mathcal{S}y}{y^T\mathcal{S}^T\mathcal{S}y} + \frac{z^T\alpha Cz}{z^T\alpha Cz} \\ &\leq 1 + \frac{\alpha}{\gamma} + 1, \quad \text{(B.33)}\end{aligned}$$

using the second part of (B.30) and the fact that $y^T\mathcal{S}^T\mathcal{S}y = y^T\mathcal{S}^T\tilde{V}_\gamma^T\tilde{V}_\gamma\mathcal{S}y$. In the same way, we can write

$$\begin{aligned}\frac{1}{\nu(x)} &\leq \frac{y^T\mathcal{C}^T V_\gamma^T(\alpha\Lambda_\gamma^{-1})V_\gamma\mathcal{C}y}{y^T\mathcal{C}^T V_\gamma^T(\alpha\Lambda_\gamma^{-1})V_\gamma\mathcal{C}y} + \frac{y^T\mathcal{C}^T V_\gamma^T V_\gamma\mathcal{C}y}{y^T\mathcal{C}^T V_\gamma^T(\alpha\Lambda_\gamma^{-1})V_\gamma\mathcal{C}y} \\ &\quad + \frac{y^T\mathcal{S}^T\mathcal{S}y}{y^T\mathcal{S}^T\tilde{V}_\gamma^T(\alpha\tilde{\Lambda}_\gamma^{-1})\tilde{V}_\gamma\mathcal{S}y} + \frac{z^T\alpha Cz}{z^T\alpha Cz} \\ &\leq 1 + \frac{\gamma}{\alpha} + \frac{\lambda_{\max}(A)}{\alpha} + 1, \quad \text{(B.34)}\end{aligned}$$

where we use both parts of (B.30) and the fact that $y^T\mathcal{S}^T\mathcal{S}y = y^T\mathcal{S}^T\tilde{V}_\gamma^T\tilde{V}_\gamma\mathcal{S}y$. The bounds (B.33) and (B.34) imply the desired result for this first case. Now, from $\mathcal{C}^T\mathcal{C} + \mathcal{S}^T\mathcal{S} = I_m$, one cannot have both $\mathcal{C}y = 0$ and $\mathcal{S}y = 0$ at the same time for a non-zero $x \in \mathbb{R}^m$. Considering these particular cases separately, we can deduce from (B.30), (B.31) and (B.32) when $\mathcal{C}y = 0$

$$\begin{aligned}\nu(x) &= \frac{y^T\mathcal{S}^T\tilde{V}_\gamma^T(\alpha\tilde{\Lambda}_\gamma^{-1})\tilde{V}_\gamma\mathcal{S}y}{y^T\mathcal{S}^T\mathcal{S}y + z^T\alpha Cz} + \frac{z^T\alpha Cz}{y^T\mathcal{S}^T\mathcal{S}y + z^T\alpha Cz} \\ &\leq \frac{\alpha}{\gamma} + 1 \\ &\leq 1 + \frac{\alpha}{\gamma} + 1,\end{aligned}$$

and

$$\begin{aligned}\frac{1}{\nu(x)} &= \frac{y^T\mathcal{S}^T\mathcal{S}y}{y^T\mathcal{S}^T\tilde{V}_\gamma^T(\alpha\tilde{\Lambda}_\gamma^{-1})\tilde{V}_\gamma\mathcal{S}y + z^T\alpha Cz} + \frac{z^T\alpha Cz}{y^T\mathcal{S}^T\tilde{V}_\gamma^T(\alpha\tilde{\Lambda}_\gamma^{-1})\tilde{V}_\gamma\mathcal{S}y + z^T\alpha Cz} \\ &\leq \frac{\lambda_{\max}(A)}{\alpha} + 1. \\ &\leq 1 + \frac{\gamma}{\alpha} + \frac{\lambda_{\max}(A)}{\alpha} + 1.\end{aligned}$$

Similarly, when $\mathcal{S}y = 0$, we decuce from (B.31) and (B.32),

$$\begin{aligned}\nu(x) &= \frac{y^T\mathcal{C}^TV_\gamma^T(\alpha\Lambda_\gamma^{-1})V_\gamma\mathcal{C}y}{y^T\mathcal{C}^TV_\gamma^T(\alpha\Lambda_\gamma^{-1} + I_p)V_\gamma\mathcal{C}y + z^T\alpha Cz} \\ &\qquad + \frac{z^T\alpha Cz}{y^T\mathcal{C}^TV_\gamma^T(\alpha\Lambda_\gamma^{-1} + I_p)V_\gamma\mathcal{C}y + z^T\alpha Cz} \\ &\leq 1 + 1 \\ &\leq 1 + \frac{\alpha}{\gamma} + 1,\end{aligned}$$

and

$$\begin{aligned}\frac{1}{\nu(x)} &= \frac{y^T\mathcal{C}^TV_\gamma^T(\alpha\Lambda_\gamma^{-1} + I_p)V_\gamma\mathcal{C}y}{y^T\mathcal{C}^TV_\gamma^T(\alpha\Lambda_\gamma^{-1})V_\gamma\mathcal{C}y + z^T\alpha Cz} \\ &\qquad + \frac{z^T\alpha Cz}{y^T\mathcal{C}^TV_\gamma^T(\alpha\Lambda_\gamma^{-1})V_\gamma\mathcal{C}y + z^T\alpha Cz} \\ &\leq 1 + \frac{\gamma}{\alpha} + 1 \\ &\leq 1 + \frac{\gamma}{\alpha} + \frac{\lambda_{\max}(A)}{\alpha} + 1.\end{aligned}$$

In both cases, the bounds in (B.33) and (B.34) are still valid, which ends the proof. □

# Abbreviations and main notations

| | | |
|---|---|---|
| $A$ | Square and symmetric matrix $\left(\in \mathbb{R}^{n\times n}\right)$ | 1 |
| $n$ | Size of the matrix $A$ | 1 |
| $B$ | Rectangular matrix with full column rank $\left(\in \mathbb{R}^{n\times m}, n \leq m\right)$ | 1 |
| $C$ | Square matrix $\left(\in \mathbb{R}^{m\times m}\right)$ | 1 |
| $m$ | Size of the matrix $C$ | 1 |
| $\mathcal{A}$ | Saddle-point matrix $\left(\in \mathbb{R}^{(n+m)\times(n+m)}\right)$ | 1 |
| KKT | Karush-Kuhn-Tucker | 1 |
| SQD | Symmetric quasi-definite | 1 |
| $\mathcal{A}_{KKT}$ | Symmetrix matrix of the Karush-Kuhn-Tucker form $\left(\in \mathbb{R}^{(n+m)\times(n+m)}\right)$ | 1 |
| $\mathcal{A}_{SQD}$ | Symmetric quasi-definite matrix $\left(\in \mathbb{R}^{(n+m)\times(n+m)}\right)$ | 2 |
| $\nabla f(x)$ | Gradient of function $f$ at point $x$ | 3 |
| $\nabla^2 f(x)$ | Hessian of function $f$ at point $x$ | 3 |
| $J(x)$ | Jacobian matrix of the constraints at point $x$ | 6 |
| $J_{\mathcal{E}}(x)$ | Jacobian matrix of the equality constraints at point $x$ | 6 |
| $J_{\mathcal{I}}(x)$ | Jacobian matrix of the inequality constraints at point $x$ | 6 |
| LICQ | Linear independency constraint qualification | 6 |
| SQP | Sequential quadratic programming | 9 |
| CG | Conjugate gradient | 25 |
| Minres | Minimal residuals | 27 |
| PCG | Preconditioned conjugate gradient | 30 |
| $S$ | Schur complement $\left(\in \mathbb{R}^{m\times m}\right)$ | 33 |
| $\mathcal{P}$ | Exact block diagonal preconditioner $\left(\in \mathbb{R}^{(n+m)\times(n+m)}\right)$ | 33 |
| $\tilde{\mathcal{P}}$ | Approximation of exact block diagonal preconditioner $\left(\in \mathbb{R}^{(n+m)\times(n+m)}\right)$ | 35 |
| $\tilde{A}$ | Approximation of matrix $A$ $\left(\in \mathbb{R}^{n\times n}\right)$ | 35 |
| $\tilde{S}$ | Approximation of schur complement $\left(\in \mathbb{R}^{m\times m}\right)$ | 35 |

| | | |
|---|---|---|
| $\mathcal{P}_{\mathcal{GGV}}$ | Golub-Greif-Varah preconditioner $\left(\in \mathbb{R}^{(n+m)\times(n+m)}\right)$ | 35 |
| $\mathcal{P}_c$ | Constraint preconditioner $\left(\in \mathbb{R}^{(n+m)\times(n+m)}\right)$ | 44 |
| $\Lambda_\gamma$ | Diagonal matrix containing the $p$ eigenvalues less than $\gamma$ $\left(\in \mathbb{R}^{p\times p}\right)$ | 46 |
| $\gamma$ | Positive number $\in [\lambda_{\min}(A), \lambda_{\max}(A)]$ $\left(\in \mathbb{R}^{+}\right)$ | 46 |
| $\lambda_{\min}(A)$ | The largest eigenvalue of $A$ | 46 |
| $\lambda_{\max}(A)$ | The smallest eigenvalue of $A$ | 46 |
| $\tilde{\Lambda}_\gamma$ | Diagonal matrix containing the $n-p$ eigenvalues more than $\gamma$ $\left(\in \mathbb{R}^{(n-p)\times(n-p)}\right)$ | 46 |
| $U_\gamma$ | Rectangular matrix such that the columns are the orthonormal sets of eigenvectors corresponding to $\Lambda_\gamma$ $\left(\in \mathbb{R}^{n\times p}\right)$ | 46 |
| $\tilde{U}_\gamma$ | Rectangular matrix such that the columns are the orthonormal sets of eigenvectors corresponding to $\tilde{\Lambda}_\gamma$ $\left(\in \mathbb{R}^{n\times(n-p)}\right)$ | 46 |
| $\alpha$ | Estimate of the average of the eigenvalues in $\tilde{\Lambda}_\gamma$ $\left(\in \mathbb{R}^{+}\right)$ | 46 |
| SLRU | Spectral low rank update | 46 |
| $A_\gamma^{-1}$ | SLRU approximation of the inverse of matrix $A$ $\left(\in \mathbb{R}^{n\times n}\right)$ | 46 |
| LMP | Limited-memory preconditioner | 48 |
| $S_\gamma$ | Approximation of the Schur complement $S$ $\left(\in \mathbb{R}^{m\times m}\right)$ | 50 |
| $S_\gamma^{-1}$ | Approximation of the inverse of the Schur complement $S$ $\left(\in \mathbb{R}^{m\times m}\right)$ | 50 |
| LSC | Least-squares commutator | 83 |

# Bibliography


A. Altman and J. Gondzio. Regularized symmetric indefinite systems in interior point methods for linear and quadratic optimization. Working papers, Ecole des Hautes Etudes Commerciales, Universite de Geneve-, 1998.

A. Battermann and M. Heinkenschloss. Preconditioners for Karush-Kuhn-Tucker systems arising in the optimal control of distributed systems. *Optimal Control of Partial Differential Equations*, pp. 17–37, 1997.

A. Battermann and E. W. Sachs. Block preconditioners for KKT systems arising in PDE-governed optimal control problems. *Fast Solution of Discretized Optimization Problems*, pp. 1–18, 2001.

M. Benzi. Preconditioning techniques for large linear systems: a survey. *J. Comput. Phys.*, **182**, 418–477, 2002.

M. Benzi and A. Wathen. Some preconditioning techniques for saddle point problems. *in* W. Schilders, H. van der Vorst and J. Rommes, eds, 'Model Order Reduction: Theory, Research Aspects and Applications', Vol. 13 of *Mathematics in Industry*, pp. 195–211. Springer Berlin Heidelberg, 2008.

M. Benzi, G.H. Golub, and J. Liesen. Numerical solution of saddle point problems. *Acta Numerica*, **14**, 1–137, 2005.

Å. Björck. *Numerical Methods for Least Squares Problems*. SIAM, Philadelphia, USA, 1996.

B. Carpentieri, I.S. Duff, and L. Giraud. A class of spectral two-level preconditioners. *SIAM J. Sci. Comput.*, **25(2)**, 749–765, 2003.

A.R. Conn, N.I.M. Gould, and Ph.L. Toint. *Trust-Region Methods*. Number 01 in MPS-SIAM Series on Optimization. SIAM, Philadelphia, USA, 2000.

H.S. Dollar and A.J. Wathen. Incomplete factorization constraint preconditioners for saddle-point matrices. Technical Report NA-04/01, Numerical Analysis Group, Oxford University, 2004.

H. C. Elman, A. Ramage, and D. J. Silvester. Ifiss: a matlab toolbox for modelling incompressible flow. *SIAM Journal in Numerical Analysis*, **40**(40), 254–281, 2002.

H.C. Elman, D.J. Silvester, and A.J. Wathen. *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics*. Oxford University Press, Oxford, 2005.

M.C. Ferris. Finite termination of the proximal point algorithm. *Mathematical Programming*, **26**, 359–366, 1991.

A.V. Fiacco and G.P. McCormick. *Nonlinear programming: Sequential Unconstrained Minimization Techniques*. John Willey & Sons, 1968.

B. Fischer. *Polynomial Based Iteration Methods for Symmetric Linear Systems*. SIAM, Philadelphia, 2011.

B. Fischer, A. Ramage, D.J. Silvester, and A.J. Wathen. Minimum residual methods for augmented systems. *BIT*, **38(3)**, 527–543, 1998.

A. Forsgren, P.E. Gill, and M.H. Wright. Interior methods for nonlinear optimization. *SIAM Review*, **44**, 525–597, 2002.

A. Forsgren, P. E. Gill, and E. Wong. Primal and dual active-set methods for convex quadratic programming. *Mathematical Programming*, pp. 1–40, 2015.

M. P. Friedlander and D. Orban. A primal-dual regularized interior-point method for convex quadratic programs. *Mathematical Programming Computation*, **4**, 71–107, 2012.

K.R. Frisch. The logarithmic potential method for conex programming. Manuscript, University Institute of Economics, Oslo, 1955.

P.E. Gill, W. Murray, and M.H. Wright. *Practical Optimization*. Academic Press, London, 1981.

L. Giraud, D. Ruiz, and A. Touhami. A comparative study of iterative solvers exploiting spectral information for SPD systems. *SIAM J. Sci. Comput.*, **27(5)**, 1760–1786, 2006.

G.H. Golub and Ch. Greif. On solving block-structured indefinite linear systems. *SIAM J. Sci. Comput.*, **24(6)**, 2076–2092, 2003.

G.H. Golub and Ch.F. Van Loan. *Matrix Computations*. Johns Hopkins, 1996.

G.H. Golub and Ch.F. Van Loan. *Matrix Computations*. Johns Hopkins, 2013.

G.H. Golub, Ch. Greif, and J.M. Varah. An algebraic analysis of a block diagonal preconditioner for saddle point systems. *SIAM J. Matrix Anal. Appl.*, **27(3)**, 779–792, 2006.

G.H. Golub, D. Ruiz, and A. Touhami. A hybrid approach combining Chebyshev filter and conjugate gradient for solving linear systems with multiple right-hand sides. *SIAM J. Matrix Anal. Appl.*, **29(3)**, 774–795, 2007.

N.I.M. Gould and V. Simoncini. Spectral analysis of saddle point matrices with indefinite leading blocks. *SIAM J. Matrix Anal. Appl.*, **31(3)**, 1152–1171, 2009.

N.I.M. Gould, M.E. Hribar, and Ph.L. Toint. On the solution of equality constrained quadratic problems arising in optimization. *SIAM J. Sci. Comput.*, **23**, 1375–1394, 2001*a*.

N.I.M. Gould, B. Orban, and Ph. L. Toint. Cuter (and sifdec), a constrained and unconstrained testing environment, revisited. Technical Report TR/-PA/01/04, Optimization Technology Center, Northwestern University, Evanston, Illinois, USA, 2001*b*.

N. Gould, D. Orban, and P. Toint. Numerical methods for large-scale nonlinear optimization. *Acta Numerica*, **14**, 299–361, 2005.

A. Greenbaum. *Iterative Methods for Solving Linear Systems*. SIAM, Philadelphia, 1997.

C. Greif and D. Schötzau. Preconditioners dor saddle point linear systems with highly singular $(1, 1)$ blocks. *Electronic Transactions on Numerical Analysis*, **22**, 114–121, 2006.

L. A. Hageman and D. M. Young. *Applied Iterative Methods*. Academic Press, New York and London, 1981.

M.R. Hestenes. Multiplier and gradient methods. *Journal of Optimization Theory and Applications*, **4(5)**, 303–320, 1969.

M.R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of the National Bureau of Standards*, **49**, 409–436, 1952.

N. J. Higham and S. H. Cheng. Modifying the inertia of matrices arising in optimization. *Linear Algebra and its Applications*, **275–276**, 261–279, 1998.

J.B. Hiriart-Urruty. *L'optimisation*. Presse Universitaire de France, Paris, 1996.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, New York, 1985.

W. Karush. Minima of functions of several variables with inequalities as side conditions. Master's thesis, Department of Mathematics, University of Chicago, Illinois, USA, 1939.

C. Keller, N.I.M. Gould, and A.J. Wathen. Constraint preconditioning for indefinite linear systems. *SIAM J. Matrix Anal. Appl.*, **21(4)**, 1300–1317, 2000.

H.W. Kuhn and A.W. Tucker. Nonlinear programming. *in* 'Proceedings of the second Berkeley symposium on mathematical statistics and probability', California, USA, 1950. University of Berkeley Press.

P. Lancaster and M. Tismenetsky. *The Theory of Matrices.* Academic Press, San Diego,California, USA, second edn, 1985.

C. Lanczos. Solution of systems of linear equations by minimized iterations. *Journal of Research of National Bureau of Standards*, **49(1)**, 33–53, 1952.

D.G. Luenberger. *Introduction to Linear and Nonlinear Programming.* Addison-Wesley Publishing Company, Philippines, 1973.

G. Meurant. *The Lanczos and conjugate gradient algorithms from theory to finite precision computations.* SIAM, Philadelphia, 2006.

G. Meurant and Z. Strakoš. The Lanczos and conjugate gradient algorithms in finite precision arithmetic. *Acta Numerica*, **15**, 471–542, 2006.

C. D. Meyer. *Matrix Analysis and Applied Linear Algebra.* SIAM, Philadelphia, 2000.

M.F. Murphy, G.H. Golub, and A.J. Wathen. A note on preconditioning for indefinite linear systems. *SIAM J. Sci. Comput.*, **21(6)**, 1969–1972, 2000.

J. Nocedal and S.J. Wright. *Numerical Optimization.* Series in Operation Research. Springer Verlag, Heidelberg, Berlin, New York, second edition, 2006.

M.A. Olshanskii and V. Simoncini. Acquired clustering properties and solution of certain sadlle point systems. *SIAM J. Matrix Anal. Appl.*, **31(5)**, 2754–2768, 2010.

C.C. Paige and M.A. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM J. Numer. Anal.*, **12**, 617–629, 1975.

C.C. Paige and M.A. Saunders. Towards a generalized singular value decomposition. *SIAM J. Numer. Anal.*, **18(3)**, 398–405, 1981.

I. Perugia, V. Simoncini, and M. Arioli. Linear algebra methods in a mixed approximation of magnetostatic problems. *SIAM J. Sci. Comput.*, **21(3)**, 1085–1101, 1999.

J. Pestana and A.J. Wathen. Natural preconditioners for saddle point systems. Technical report, The Mathematical Institute, University of Oxford, UK, 2014.

M. J. D. Powell. A method for nonlinear constraints in minimization problems. *in* R. Fletcher, ed., 'Optimization', pp. 283–298. Academic Press, New York, 1969.

A. van der Sluis and H.A. van der Vorst. The rate of convergence of conjugate gradients. *Numerische Mathematik*, **48(5)**, 543–560, 1986.

H. A. van der Vorst. *Iterative krylov Methods for Large Linear Systems*. Cambridge University Press, 2003.

R. T. Rockafellar. Augmented lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of Operations Research*, **1(2)**, 97–116, 1976.

T. Rusten and R. Winther. A preconditioned iterative method for saddlepoint problems. *SIAM J. Matrix Anal. Appl.*, **13(3)**, 887–904, 1992.

Y. Saad. *Iterative Methods for Sparse Linear Systems: Second Edition*. SIAM, Philadelphia, 2003.

Y. Saad. *Numerical Methods for Large Eigenvalue Problems*. SIAM, Philadelphia, 2011.

D. Silvester and A. Wathen. Fast iterative solution of stabilised stokes systems. part II: Using general block preconditioners. *SIAM J. Numer. Anal.*, **31(5)**, 1352–1367, 1994.

J. Tshimanga. *On a class of limited memory preconditioners for large scale nonlinear least-squares problems*. PhD thesis, Department of Mathematics, FUNDP, Belgium, 2007.

R.J. Vanderbei. Symmetric quasi-definite matrices. *SIAM Journal on Optimization*, **5(1)**, 100–113, 1995.

R.B. Wilson. *A simplicial algorithm for concave programming*. PhD thesis, Harvard University, Massachussetts, USA, 1963.

S.J. Wright. *Primal-Dual Interior-Point Methods*. SIAM, Philadelphia, USA, 1997.

F. Zhang. *The Schur complement and its applications*. Springer, 2010.